UNIVERSIDAD DE CÓRDOBA

Programa de Doctorado en
Computación Avanzada, Energía y Plasmas
Departamento de Informática y Análisis Numérico

Text Mining y Medicina: Una aproximación a la detección temprana de enfermedades

Text Mining and Medicine: An approach to early detection of diseases

MEMORIA DE TESIS PRESENTADA POR

Mª del Carmen Luque Guzmán

DIRECTORES
Dr. Sebastián Ventura Soto
Dr. José María Luna Ariza

Córdoba Junio de 2020

TITULO: *TEXT MINING Y MEDICINA: UNA APROXIMACIÓN A LA DETECCIÓN TEMPRANA DE ENFERMEDADES*

AUTOR: *María del Carmen Luque Guzmán*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es

TÍTULO DE LA TESIS: Text Mining y Medicina: Una aproximación a la detección temprana de enfermedades

DOCTORANDO/A: Mª del Carmen Luque Guzmán

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

En la presente tesis doctoral se ha realizado un estudio detallado del estado del arte en lo referente a Text Mining y medicina. Dicho estudio ha permitido publicar un artículo de revisión en una revista de alto impacto (*Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*). Con este análisis del estado del arte, se diseñó una herramienta, llamada MiNerDoc, cuyo objetivo es la de facilitar la toma de decisiones clínicas. Esta herramienta ha permitido, entre otras funcionalides, detectar factores de riesgo o eventos clínicos de interés e inferir automáticamente códigos de diagnósticos normalizados utilizando información textual como entrada. Una primera versión de esta herramienta fue presentada en el congreso internacional CMBS 2019, teniendo una gran aceptación por la comunidad. El trabajo fue propuesto para un número especial en la revista Computational Intelligence. La extensión de dicho trabajo fue publicada posteriormente en dicha revista, en el año 2020.

Con todo lo anterior, los directores de esta tesis doctoral consideran que el informe de la presente tesis doctoral es muy favorable.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 16 de Junio de 2020

Firma del/de los director/es

VENTURA SOTO SEBASTIAN EMILIO - 30510000V

Firmado digitalmente por VENTURA SOTO SEBASTIAN EMILIO - 30510000V
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-30510000V, givenName=SEBASTIAN EMILIO, sn=VENTURA SOTO, cn=VENTURA SOTO SEBASTIAN EMILIO - 30510000V
Fecha: 2020.06.17 00:02:28 +02'00'

Fdo.: Sebastián Ventura Soto

LUNA ARIZA JOSE MARIA - 30979388K

Firmado digitalmente por LUNA ARIZA JOSE MARIA - 30979388K
Fecha: 2020.06.16 22:07:04 +02'00'

Fdo.: José María Luna Ariza

La memoria titulada "*Text Mining y Medicina: Una aproximación a la detección temprana de enfermedades*" , que presenta Mª del Carmen Luque Guzmán para optar al grado de Doctor, ha sido realizada dentro del programa de doctorado "*Computación avanzada, energía y plasmas*" del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba, bajo la dirección de los doctores Sebastián Ventura Soto y José María Luna Ariza cumpliendo, en su opinión, los requisitos exigidos a este tipo de trabajos.

Córdoba, Junio de 2020

El Doctorando

Firmado por LUQUE GUZMAN MARIA DEL CARMEN - 30537643Z el día 16/06/2020 con un certificado emitido por AC FNMT Usuarios

Fdo: Mª del Carmen Luque Guzmán

El Director

VENTURA SOTO SEBASTIAN EMILIO - 30510000V

Firmado digitalmente por VENTURA SOTO SEBASTIAN EMILIO - 30510000V
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-30510000V, givenName=SEBASTIAN EMILIO, sn=VENTURA SOTO, cn=VENTURA SOTO SEBASTIAN EMILIO - 30510000V
Fecha: 2020.06.17 00:03:11 +02'00'

Fdo: Dr. Sebastián Ventura Soto

El Director

LUNA ARIZA JOSE MARIA - 30979388K

Firmado digitalmente por LUNA ARIZA JOSE MARIA - 30979388K
Fecha: 2020.06.16 22:07:32 +02'00'

Fdo: Dr. José María Luna Ariza

"Las cosas nunca son como a primera vista las figuramos, y así ocurre que cuando empezamos a verlas de cerca, cuando empezamos a trabajar sobre ellas, nos presentan tan raros y hasta tan desconocidos aspectos, que de la primera idea no nos dejan a veces ni el recuerdo; tal pasa con las caras que nos imaginamos, con los pueblos que vamos a conocer, que nos los hacemos de tal o de cual forma en la cabeza, para olvidarnos repentinamente ante la vista de lo verdadero. Esto es lo que me ocurrió con este papeleo, que si al principio creí que en ocho días lo despacharía, hoy -al cabo de ciento veinte- me sonrío no más que de pensar en mi inocencia".

--Camilo José Cela
La Familia de Pascual Duarte

Agradecimientos

Es difícil resumir en breves palabras mi sentimiento de gratitud a todas las personas que han hecho posible que mi sueño se convirtiera en una realidad, que han podido despertar en mí nuevas ilusiones y que me han enseñado que pueden existir otras formas de pensar y actuar.

En primer lugar quiero acordarme de mis directores de tesis. Del Dr. Sebastián Ventura no puedo decir, como otros doctorandos, que ha sido como un padre para mí sólo por una ligera cuestión de edad, pero sí puedo decir que sin su ayuda nunca hubiera llegado hasta la meta, gracias por tu continuo ánimo, humildad y generosidad. Al Dr. José María Luna le estaré siempre eternamente agradecida, por su dedicación, sus incansables revisiones y por estar siempre apoyándome en los momentos complicados. Los dos me han demostrado que todavía existen personas con grandes valores.

En segundo lugar quiero agradecer la gran ayuda prestada por la Dra. Miñarro del Moral, de la cual he intentado absorber un poquito de su gran sabiduría. No puedo olvidarme de los compañeros del grupo de investigación KDIS, a todos muchas gracias por vuestro ayuda. Especialmente quiero acordarme de José María Moyano y Oscar Reyes, su apoyo en las primeras fases fueron cruciales, siempre han estado disponibles cuando los he necesitado. Además de ser excelentes compañeros me han demostrado que son excelentes personas.

En especial, quiero dedicar todos estos años de estudio y dedicación a mis padres, a los que pronto podré dedicar todo el tiempo que ahora necesitan, a mis hermanas, que han aguantado mis momentos de bajón y me han sabido animar en los días más críticos y a dos de mis seres más queridos, mi marido y mi hijo, a los que únicamente tengo que darles las gracias y pedirles perdón porque para llegar hasta aquí han sido inevitables mis grandes ausencias en momentos importantes.

Por último, quería dejar por escrito una frase que espero pueda servir de ayuda para algunos doctorandos que lleguen detrás: "*Nunca digas no puedo, no hay género, edad, ni excusa que te impida cumplir tu sueño. Todo depende de ti*."

Resumen

El futuro próximo de los servicios sanitarios vendrá marcado por el envejecimiento de la población y la cronicidad de las enfermedades. Junto a los cambios demográficos y sociales, se está produciendo un claro aumento de la frecuentación en los distintos servicios de atención primaria y especializada y, por supuesto, todo esto se traduce en un fuerte incremento del gasto sanitario. Todo este problemático contexto hace que las instituciones sanitarias se marquen como principales objetivos la priorización de la prevención, el control de los factores de riesgo y la detección precoz de enfermedades. Para apoyar la prevención primaria es muy importante que el profesional sanitario tenga todos los medios disponibles a su alcance para extraer conocimiento de su principal fuente de información que es la historia clínica informatizada del paciente. Así, el profesional sanitario debería disponer de herramientas que permitan conocer e interrelacionar eventos clínicos de interés, alertar sobre la aparición de futuros riesgos para la salud o pronosticar el posible desarrollo de una enfermedad. Sin embargo, el esfuerzo, tiempo y coste que supondría extraer este conocimiento de la simple lectura de los múltiples informes clínicos contenidos en la historia de un paciente (escritos en su mayoría en lenguaje natural), sería incalculable e imposible de asumir por la mayoría de los profesionales sanitarios en la clínica diaria.

Hasta el momento, los sistemas de información existentes en la mayoría de instituciones sanitarias sólo han sido utilizados como sistemas de almacenaje de información, es decir sistemas que recopilan y almacenan toda la información asistencial generada en la interacción médico-paciente, pero todavía no se ha dado el paso de convertir estos grandes “almacenes de información” en “fuentes de conocimiento” que aporten valor para facilitar y apoyar la toma de decisiones clínicas.

Sin embargo, el reto de automatizar este proceso, transformar almacenes de información en fuentes de conocimiento, no es una tarea trivial. Se estima que en un complejo hospitalario regional se pueden generar al año más de 3 millones de documentos clínicos, el 80% de esta documentación clínica contiene información no

estructurada, una de la más destacable es la información textual. Hasta ahora la información clínica textual ha sido prácticamente ignorada por la mayoría de las instituciones sanitarias debido a la gran complejidad en su explotación para generar valor de su contenido. La principal fuente de conocimiento contenida en la historia clínica electrónica, que es la narrativa clínica textual, es en la práctica altamente desaprovechada. A la dificultad de las organizaciones sanitarias para obtener valor del texto, con las herramientas de análisis hasta ahora utilizadas, se suman las peculiares características que posee la terminología clínica donde prima: una alta ambigüedad y complejidad del vocabulario, la narrativa textual libre, una escasa normalización terminológica y un uso excesivo de acrónimos y negaciones.

En este complejo marco y ante la creciente necesidad de adquirir conocimiento para apoyar el proceso de prevención y toma de decisiones clínicas, se hace imprescindible el uso de Sistemas Inteligentes que ayuden a extraer el valor encerrado en el contenido textual de los múltiples documentos que integran la historia clínica electrónica. Pero a pesar de esta acuciante necesidad, actualmente existen muy pocos sistemas reales que extraigan conocimiento del texto clínico para facilitar el trabajo diario al profesional sanitario en tareas arduas y complejas como la detección de factores de riesgo o la predicción diagnóstica. En la actualidad, para abordar la problemática de extraer valor del texto clínico, en el entorno de la medicina computacional, disponemos de las técnicas avanzadas que nos proporciona la disciplina de la Minería de Textos (MT). Esta disciplina puede definirse como un área orientada a la identificación y extracción de nuevo conocimiento adquirido a partir de información textual, es un campo multidisciplinar que puede integrar técnicas de otras disciplinas como el Procesamiento del Lenguaje Natural (PLN) o Aprendizaje Automático (AA).

En este sentido, abordamos esta tesis doctoral con un análisis exhaustivo y pormenorizado del estado del arte sobre la disciplina de la MT en el ámbito de la Medicina, recogiendo los métodos, técnicas, tareas, recursos y tendencias más destacadas en la literatura. De esta amplia revisión se detecta que en la práctica los sistemas existentes para apoyar el proceso de toma de decisiones clínicas basados en información clínica textual son escasos y generalmente resuelven una única tarea

principal centrándose en un área específica de conocimiento y siendo desarrollados para dominios muy específicos difícilmente reproducibles en otros entornos. Ante las problemáticas observadas en los sistemas de MT existentes y las necesidades de las instituciones sanitarias, se propone la creación de un novedoso sistema, denominado MiNerDoc, que permita apoyar la toma de decisiones clínicas en base a una combinación de técnicas de la disciplina de la MT, junto con el enriquecimiento terminológico y semántico proporcionado por la herramienta MetaMap y el metathesaurus UMLS, recursos que aportan características esenciales en el dominio médico. MiNerDoc permite, entre otras funcionalidades, detectar factores de riesgo o eventos clínicos de interés e inferir automáticamente códigos normalizados de diagnósticos tomando como fuente exclusiva la información textual contenida en informes clínicos, en definitiva, permite llevar a cabo tareas complejas que facilitan y apoyan la labor del profesional sanitario en la prevención primaria y la toma de decisiones clínicas. El sistema de MT propuesto ha sido evaluado en base a un amplio análisis experimental, los resultados demostraron la efectividad y viabilidad del sistema propuesto y verificaron el prometedor rendimiento de MiNerDoc en las dos tareas evaluadas, reconocimiento de entidades médicas y clasificación diagnóstica multietiqueta.

Abstract

The near future of health services will be marked by the ageing of the population and the chronicity of diseases. Together with the demographic and social changes, there is a clear increase in the number of people attending both primary and specialized care services, and, of course, all this produces a sharp increase in healthcare expenditure. All this context makes health institutions to set a series of main objectives: prioritization of prevention, control of risk factors and early detection of diseases. To support primary prevention, it is important that health professionals have all the available means at their disposal to extract knowledge from main sources of information, that is, the patient's electronic health records. Thus, health professionals should have tools that allow them to know and interrelate clinical events of interest, receive alerts about upcoming health risks or predict the development of a disease. However, the effort, time and cost required to extract this knowledge by just reading of the multiple clinical reports belonging to a patient's history (mostly written in natural language), are incalculable and hardly affordable for most health professionals in the daily clinic practice.

Until now, the existing information systems in most health institutions have only been used as information storage systems, that is, systems that collect and store any healthcare information generated in the practitioner-patient interaction. By now, the step of transforming such raw data into useful "knowledge" that eases and supports the final clinical decision-making process has not been applied yet. Nevertheless, such challenge of transforming raw data into knowledge is not trivial. It is estimated that in a regional hospital more than 3 million clinical documents can be generated per year, 80% of them contain unstructured or textual information. Up to now, textual clinical information has been practically ignored by most health institutions mainly due to the arduous process required to take advantage of the content of such vast amount of data. Thus, the main source of knowledge contained in the electronic medical records, which is in textual clinical narrative, is practically untapped. Additionally to the difficulty of the health organizations to obtain value from the text by using traditional tools, the peculiar characteristics of the clinical terminology is an added problem: high ambiguity and

complexity of the vocabulary, free textual narrative, a poor terminological standardization and an overuse of acronyms and negations.

In this complex framework and in view of the growing need to acquire knowledge to support the decision-making process, it is essential to use Intelligent Systems that help to extract the value from textual documents**.** Currently, there are very few real systems able to extract knowledge from clinical texts and to really ease the daily work of healthcare professionals in complex tasks such as risk factor detection or diagnostic prediction. In recent years, to face these problems up, there are a number of advanced techniques provided by the Text Mining (TM) discipline. TM might be defined as an area focused on the identification and extraction of new knowledge from textual information, and it is seen as a multidisciplinary field gathering techniques from other disciplines such as Natural Language Processing (NLP) and Machine Learning (ML).

In this sense, this doctoral Thesis first provides an exhaustive and detailed analysis of the state-of-the-art on the TM discipline in Medicine. This analysis includes the most outstanding methods, techniques, tasks, resources and trends in the field. As a result, this review revealed that the existing systems to support the clinical decision-making process by applying a textual clinical information are scarce, and they generally perform a single task on a specific area of knowledge and for very specific domains hardly applied to problems on different environments. In this regard, this Thesis proposes the development of a new system, called MiNerDoc, to support clinical decision-making by applying a combination of techniques from the TM discipline, along with the terminological and semantic enrichment provided by the MetaMap tool and the UMLS metathesaurus. MiNerDoc allows, among other functionalities, the detection of risk factors or clinical events of interest and automatic inference of standardized diagnostic codes based on the textual information included in clinical reports. The proposed TM system has been evaluated based on an extensive experimental study and the results have demonstrated the effectiveness and viability of such system in two tasks, recognition of medical entities and multi-label diagnostic classification.

Índice de Contenidos

Índice de Figuras

Índice de Tablas

Lista de Acrónimos

AA	Aprendizaje Automático
IA	Inteligencia Artificial
ADR	Efectos adversos a medicamentos
BoW	Bag of Word
BR	Binary Relevance
CAD	Clasificación automática de documentos
CC	Classifier Chain
CIE	Clasificación Internacional de Enfermedades
CLAMP	Clinical Language Annotation, Modeling, and Processing Toolkit
CLEF	Cross Language Evaluation Forum
CliNER	Clinical Named Entity Recognition system
CRF	Conditional Random Fields
CUI	Concept Unique Identifiers
dCSE	Diagnostic Classification with Semantic Enrichement
ECC	Ensemble Classifier Chain
EI	Extracción de Información
EPS	Ensemble of Pruned Set
HOMER	Hierarchy of multi-label classifiers
IBLR	instance-based learning by logistic regression
IDF	Inverse Document Frequency
LP	Label PowerSet
LSA	Análisis Semántico Latente
MER	Medical Entity Recognition
MeSH	Medical Subject Headings
MIMIC	Medical Information Mart for Intensive Care
ML-kNN	Multi-Label K-Nearest Neighbors
MLL	Multilabel Learning
MMI	MetaMap IndexingAlgorithm
MMTx	MetaMap Transfer

MT	Minería de Textos
MUC	Message Understanding Conference
NER	Named Entity Recognition
PLN	Procesamiento del Lenguaje Natural
POST	Part-of-speech-Tagging
PS	Pruned Sets
RAkEL	RAmdom k labELsets
RI	Recuperación de información
SEI	Sistema de extracción de la información
SMO	Sequential Minimal Optimization
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency-Inverse document frequency
TREC	Text Retrieval Conference
UMLS	Unified Medical Language System
VSM	Vector Space Model
WSD	Word-Sense Disambiguation server

1

Introducción

1.1. Motivación

Uno de los principales retos y desafíos de la Sanidad, en los últimos años, se centra en facilitar un marco adecuado para la promoción de la salud y la prevención de enfermedades, hecho que se refleja en el fomento de planes y líneas estratégicas orientadas a promover la prevención primaria y el bienestar de la población. Son evidentes las numerosas ventajas que aportan tanto la prevención como el diagnóstico precoz de enfermedades, no solamente desde la perspectiva del paciente, con un mejora de la calidad de vida, sino desde el punto de vista institucional, con una importante disminución del coste asistencial.

Es importante que la investigación se vuelque en la prevención para mejorar el pronóstico de enfermedades como el cáncer o las enfermedades cardiovasculares, en las que se prevé un aumento considerable de incidencia en todo el mundo. Según la Sociedad Española de Oncología Médica, durante el año 2017 los nuevos casos estimados de cáncer en España fueron 228.482, y se estima que en 2035 se diagnosticarán 315.413 casos nuevos de cáncer[1]. El Centro de Estudios Económicos y

[1] https://seom.org/publicaciones/el-cancer-en-espanyacom

Empresariales prevé que el coste de las enfermedades cardiovasculares ascenderá en 2020 a 8.800 millones de euros en España[2]. Estos son pequeños ejemplos que confirman que la inversión en prevención es la mejor vía de paliar este previsible aumento de mortalidad y gasto asistencial que se vaticina en nuestro futuro más próximo.

Para apoyar la prevención primaria es fundamental que el profesional sanitario tenga todos los medios disponibles a su alcance para extraer conocimiento de su principal fuente de información que es la historia clínica del paciente. El clínico debería disponer de herramientas que permitan descubrir e interrelacionar eventos de interés, alertar sobre la aparición de factores de riesgos o ayudar a pronosticar el desarrollo de una posible enfermedad. Si bien es cierto que el médico tiene acceso a toda la historia clínica del paciente en cada acto clínico, también es cierto que el esfuerzo, tiempo y coste que supondría extraer este conocimiento de la simple lectura de los múltiples informes clínicos, escritos en su mayoría en lenguaje natural, sería incalculable e imposible de asumir por la mayoría de los clínicos en su actividad diaria.

Además, la terminología clínica es muy diferente a las que se genera en cualquier otro ámbito debido fundamentalmente a las siguientes peculiaridades:

- Alta ambigüedad y complejidad del vocabulario.
- Escasa normalización terminológica.
- Utilización de frases no gramaticales y cortas.
- Abundancia de acrónimos.
- Prima la narrativa textual libre y no estructurada.

Aunque la informatización plena de los procesos sanitarios está presente en la mayoría de centros sanitarios, la tarea de transformar la información clínica textual en conocimiento no es una labor fácil ni trivial. Es evidente que el gran volumen de

[2] https://www.vademecum.es/noticia-140828-el+coste+de+las+enfermedades+cardiovasculares+ascender+aacute++en+2020+a+8.800+millones+de+euros+en+espa+ntilde+a_8345

información sanitaria textual no estructurada junto con la complejidad y diversidad de la terminología médica hace mucho más difícil la tarea de extracción del conocimiento.
La accesibilidad real que actualmente poseen los sanitarios al conocimiento inmerso en los datos de salud de un paciente, está bastante alejada del marco ideal para obtener una información que pueda ser útil como herramienta de análisis o de apoyo a la toma de decisiones médicas. En muchos casos la informatización de la historia clínica ha servido para "compartir" información entre todos los profesionales que intervienen en el proceso asistencial del paciente, desde los médicos de atención primaria, hasta los especialistas, radiólogos, etc. Pero todavía no se ha invertido lo necesario en hacer que esta información sea válida para la extracción real de conocimiento, donde se puedan realizar comparativas de estudios o retroalimentaciones que ayuden a los profesionales a la toma de decisiones clínicas y, en definitiva, para que aumente la calidad asistencial en la que todos saldremos beneficiados.

El principal problema al que se enfrenta actualmente el profesional sanitario no es encontrar información, existe sobreabundancia de ella, sino seleccionar la más relevante y de calidad. No se necesita más información sino más conocimiento.

Además de lo visto anteriormente no hay que perder de vista la complejidad de la actividad asistencial, la cual se dispara ante el abordaje del ***proceso diagnóstico*** donde, con ingentes cantidades de datos e información textual, proveniente principalmente de la historia clínica, el profesional sanitario debe ser capaz de inferir el mejor diagnóstico o diagnósticos posibles. El profesional sanitario debe buscar la información más relevante, interpretar pruebas, analíticas, informes de consultas, etc, obtener sus conclusiones e hipótesis en base a estos resultados y establecer un diagnóstico de calidad y fiable. Pero, ¿cómo *es posible compaginar todos estos complejos factores con una alta calidad asistencial y un bajo nivel de errores?.*

En este complejo marco y ante la creciente necesidad de adquirir conocimiento, en base al contenido textual, se hace imprescindible el uso de Sistemas Inteligentes que ayuden a extraer valor del texto para apoyar el proceso de toma de decisiones clínicas. Para crear estos Sistemas Inteligentes contamos con los recursos que nos proporcionan las ***disciplinas de la Minería de Textos (MT) y Aprendizaje Automático (AA).*** La disciplina de la MT podría definirse básicamente como un área orientada a la extracción de conocimiento a partir de información de contenido textual. Se trata de un conjunto de técnicas que se orientan al descubrimiento de un nuevo conocimiento que explícitamente era inexistente en las colecciones textuales de las que se partían. La disciplina del AA, considerada como una rama de la Inteligencia Artificial, tiene como objetivo la construcción de sistemas capaces de adquirir conocimiento y aprender automáticamente en base a un conjunto de datos de entrenamiento. La mayoría de los sistemas basados en MT se construyen sobre la base del Procesamiento del Lenguaje Natural (PLN) y AA, todas estas disciplinas son imprescindibles para el análisis y procesamiento de grandes colecciones textuales.

Las ventajas que proporcionaría la explotación de este conocimiento sumergido en la información textual de la historia clínica, utilizando técnicas del MT y AA, serían innumerables dentro del ámbito médico, entre otros aportes podríamos citar los siguientes: apoyo en el proceso de toma de decisiones diagnósticas, reducción del número de posibles errores médicos relacionados con interacciones entre medicamentos, identificación de conflictos entre distintos tratamientos sobre el mismo paciente, detección de alertas médicas, apoyo en la resolución y búsqueda del mejor diagnóstico/tratamiento, detección precoz, etc. Sin embargo, los sistemas existentes para apoyar el proceso de toma de decisiones basados en información clínica textual son muy escasos. Actualmente, existen pocos sistemas que extraigan valor del texto, de forma sencilla y ágil, para facilitar el trabajo al clínico en tareas arduas y complejas como la detección de alertas clínicas o la codificación diagnóstica de información clínica. En esta tesis doctoral, hemos abierto varias posibilidades para contribuir a facilitar la labor

de la prevención primaria creando para ello un novedoso sistema de MT cuyo principal objetivo es apoyar el proceso de toma de decisiones clínicas. Abordaremos, para ello, dos grandes tareas que se engloban dentro de las disciplinas de la MT y AA, como son la *detección de Entidades Médicas y la Clasificación Diagnóstica Multietiqueta,* basándonos para ello en información habitual de una historia clínica, como es el caso de los informes de alta. Nos respaldaremos, entre otros recursos, en el enriquecimiento terminológico y semántico que ofrece el metatesauro UMLS a través de la herramienta MetaMap.

En esta tesis doctoral llevaremos a cabo un estudio pormenorizado del estado del arte sobre la disciplina de la MT en el ámbito de la Medicina, analizando las tareas, técnicas, métodos, recursos y tendencias de mayor relevancia en la literatura. De este amplio análisis, y en base a las problemáticas observadas para obtener automáticamente conocimiento del texto clínico, abordaremos la creación de un sistema de MT, denominado MiNerDoc, con una orientación práctica, de uso amigable, y que facilite la toma de decisiones clínicas. MiNerDoc permitirá, entre otras funcionalidades, detectar factores de riesgo o eventos clínicos de interés e inferir automáticamente códigos de diagnósticos normalizados basándose exclusivamente en información textual, en definitiva, permitirá llevar a cabo tareas complejas que faciliten la prevención primaria.

1.2. Objetivos

A continuación detallaremos cuál es el objetivo general de esta tesis doctoral y los objetivos específicos que proponemos para alcanzar la meta central propuesta en este trabajo.

OBJETIVO GENERAL

Desarrollar y poner en práctica un modelo, basado en MT, capaz de transformar la información clínica textual en conocimiento que apoye al profesional sanitario en la toma de decisiones y la detección temprana de una enfermedad.

OBJETIVOS ESPECÍFICOS

1. Revisar, analizar y estudiar en profundidad la bibliografía existente relacionada con el objetivo de esta tesis, teniendo en cuenta conceptos tales como MT, Reconocimiento de Entidades Nombradas, PLN, AA, MetaMap, UMLS, etc.

2. Desarrollar una metodología que permita el reconocimiento y la extracción de entidades médicas desde informes clínicos de contenido textual. Se propone investigar sobre cinco tipos de entidades médicas: diagnóstico, farmacología, procedimientos, hallazgos/síntomas y localización anatómica. Dicha metodología seguirá un enfoque basado en diccionarios.

3. Desarrollar una metodología que permita realizar la tarea de clasificación diagnóstica multietiqueta y que sea capaz de predecir automáticamente una o varias categorías normalizadas de diagnóstico en base al contenido textual de informes clínicos. Llamaremos a esta metodología dCSE (*diagnostic Classification with Semantic Enrichement*).

4. Crear un novedoso sistema de MT, al que llamaremos MiNerDoc, cuyo principal objetivo sea apoyar el proceso de toma de decisiones clínicas mediante el análisis de informes clínicos textuales en inglés. Este sistema unificará, en base a las metodologías propuestas, dos tareas relevantes en el campo de la Medicina, la detección de factores de riesgo en base a la detección de entidades médicas y la predicción automática de códigos normalizados de diagnóstico (descriptores MeSH asociados a enfermedades).

5. Llevar el sistema MiNerDoc a un entorno que simule un escenario real, evaluando su funcionamiento con informes de alta de pacientes reales ingresados en una unidad de cuidados intensivos tomados de la colección MIMIC (datos con deidentificación).

1.3. Estructura

La memoria se ha organizado en los siguientes capítulos. En el Capítulo 1 se presentarán la motivación, objetivos y estructura de esta tesis doctoral. El Capítulo 2 analizará en profundidad los conceptos claves sobre la disciplina de la MT, describiremos cual es el objetivo de esta disciplina, analizaremos otras disciplinas importantes necesarias para desarrollar sistemas de MT, detallaremos las fases que componen un sistema de MT y las tareas más destacadas, como el reconocimiento de entidades nombradas y la clasificación automática de documentos. Una vez descrito el marco teórico relacionado con la MT, profundizaremos en el impacto de esta disciplina en el dominio de la Medicina, revisaremos las principales tareas de MT que han sido aplicadas en el ámbito clínico, citando las principales publicaciones y estudios relacionados con distintos ámbitos de la Medicina, realizaremos un análisis en profundidad sobre dos tareas de gran importancia en el ámbito de la medicina computacional como son el reconocimiento de entidades médicas y la clasificación diagnóstica automática (apoyándonos en las principales investigaciones relacionadas). Por último, recopilaremos un amplio número de recursos y herramientas de la MT orientadas al análisis textual en el ámbito de la Medicina.

En el Capítulo 3 nos centraremos en realizar una descripción detallada del sistema de MT propuesto, denominado MiNerDoc, cuya principal finalidad es facilitar el proceso de toma de decisiones clínicas apoyando la labor de prevención primaria. En este capítulo describiremos el propósito general de MiNerDoc, su arquitectura, los requerimientos y recursos software empleados para su construcción, la metodología aplicada para llevar a cabo las dos tareas principales que lo integran (Reconocimiento de Entidades Médicas y Clasificación Diagnóstica Automática). Por último, describiremos las principales funcionalidades de MiNerDoc como son el reconocimiento de entidades médicas, la detección de factores de riesgo, la detección de negaciones, la predicción diagnóstica, etc.

El Capítulo 4 recopila varios casos de estudios, basados en informes clínicos de la base de datos MIMIC, para mostrar el funcionamiento real de MiNerDoc y sus principales funcionalidades. Así se detallará, como el sistema propuesto permite el reconocimiento de cinco grupos distintos de entidades médicas y como a través de ellas se pueden detectar automáticamente los principales factores de riesgo o alertas clínicas de interés encontrados en un informe clínico, o como realizar la predicción diagnóstica de uno o un conjunto de informes clínicos mediante la asignación automática de códigos de diagnóstico normalizados (22 descriptores MeSH asociados con enfermedades).

En el Capítulo 5 se describirá el análisis experimental diseñado para evaluar los dos subsistemas que constituyen la aplicación MiNerDoc: reconocimiento de entidades médicas y clasificación diagnóstica automática. Cada uno de los dos subsistemas se evaluará adecuadamente a través de una amplia gama de experimentos, se detallará como se han construido el corpus semántico y los distintos *datasets* creados exclusivamente para evaluar los dos subsistemas que componen el sistema propuesto (gracias al apoyo de un médico experto en documentación clínica), se analizarán en profundidad los distintos experimentos realizados y finalmente, se realizará una discusión de los mismos y se resumirán las conclusiones de mayor interés. Por último, en el Capítulo 6 se recogerán las conclusiones finales de esta tesis doctoral y las propuestas de mejora o líneas de trabajo que deberán proponerse en el futuro.

2

MARCO TEÓRICO

En este capítulo realizaremos, en primer lugar, un análisis exhaustivo sobre la disciplina de la Minería de Textos (MT), abordando conceptos claves como su definición, disciplinas que nutren a la MT, fases más importantes de las que se compone cualquier sistema basado en MT y las tareas más destacadas donde se aplica esta disciplina. En segundo lugar, realizaremos un análisis pormenorizado del estado del arte sobre la aplicación de la MT en el ámbito de la Medicina, donde profundizaremos en dos de las tareas de mayor interés en el análisis de textos clínicos y que han sido el foco central de nuestra investigación, como son el reconocimiento de entidades médicas y la clasificación diagnóstica. Analizaremos las investigaciones más destacadas donde se han utilizado técnicas de MT dentro del ámbito de la Medicina y por último, recogeremos las principales herramientas y recursos terminológicos imprescindibles para la construcción de sistemas de MT y el análisis textual en Medicina.

2.1. Minería de textos

En las últimas décadas han surgido algunas disciplinas imprescindibles para el análisis y procesamiento de grandes colecciones de información textual. Una de las que reúne las

técnicas y enfoques más útiles para inferir información estructurada de alta calidad a partir de repositorios de documentos textuales no estructurados es la disciplina de la MT.

2.1.1. Definición

La MT es un área orientada a la extracción de conocimiento a partir de información de contenido textual basada en un conjunto de técnicas avanzadas que permiten descubrir eventos de interés, novedosos y que explícitamente eran inexistentes en las colecciones textuales de las que se partían [103, 104]. Los sistemas de MT permiten extraer información relevante desde distintas fuentes textuales para generar un nuevo conocimiento.

Generalmente, la mayoría de los sistemas basados en MT siguen una serie de procesos para obtener conocimiento (relevante, nuevo y desconocido a priori) partiendo de colecciones de datos textuales. Se parte como entrada de una información con contenido textual que debe ser sometida a una fase de preprocesamiento donde los datos no estructurados se "preparan" mediante diferentes técnicas (tokenización, *stemmin*g, etc). A continuación, le sigue una segunda fase donde la información se somete a un modelo de representación adecuado donde es transformada para poder ser interpretada. En la última fase, llamada fase de descubrimiento, se extrae el verdadero valor y conocimiento del texto de partida mediante la aplicación de ciertos métodos y técnicas (clasificación, agrupación, etc).

2.1.2. Disciplinas relacionadas con la MT

Existen múltiples disciplinas que nutren e influencian a la MT, algunas de las más importantes son el Procesamiento del Lenguaje Natural (PLN) [1], Extracción de Información (EI) [15] y Aprendizaje Automático (AA) [27] (Figura 2.1). Estas áreas son la

Figura 2.1. Principales disciplinas que nutren a la MT

base sobre las que se construyen la mayoría de los sistemas de MT y sus técnicas son imprescindibles para llevar a cabo el análisis y procesamiento de grandes colecciones textuales.

Procesamiento del Lenguaje Natural. La disciplina del PLN [1] nace en la década de los 60 como un área independiente de la Inteligencia Artificial (IA) y la Lingüística Computacional. El objetivo original de esta disciplina era estudiar los problemas derivados de la comprensión automática del lenguaje natural.

Algunos autores definen el PLN como una parte esencial de la IA que es capaz de formular mecanismos computacionales que facilitan la interrelación hombre-máquina [1], otros autores la definen como una disciplina que entiende la habilidad de la "máquina" para procesar y entender la información comunicada [2].

Uno de los primeros sistemas basados en PLN fue el denominado BASEBALL, desarrollado en la década de los 60, un interfaz orientado a la comprensión del lenguaje natural. En los años 70 se amplía el área de acción hacia otros campos como la enseñanza asistida por ordenador, comprensión del lenguaje, automatización de tareas, etc. Gracias al avance de la IA se pudo desarrollar el primer sistema de pregunta-respuesta basado en lenguaje natural.

Algunas de las principales aplicaciones del PLN son [3-5]: traducción automática, interfaces humano-computadora, educación asistida por ordenador, tutores inteligentes, sistemas de búsqueda de respuestas, síntesis de voz, reconocimiento del habla, análisis de sentimientos, minería de opiniones, etc.

El lenguaje natural posee una serie de complejas características que pueden hacer, si no se tratan convenientemente, que en muchas ocasiones disminuya la efectividad de los sistemas basados en PLN. Algunos de los principales problemas a los que se enfrentan la disciplina del PLN son:

- La **anáfora** se puede definir brevemente como el término empleado para hacer referencia a algo que anteriormente ya fue mencionado. Aplicando técnicas de PLN se puede paliar el problema de la resolución de las anáforas [6].

- Otro de los escollos que tienen que salvar los sistemas PLN es la resolución de la **ambigüedad**. En el lenguaje natural nos encontramos muy frecuentemente expresiones que pueden tener varios significados diferentes según el contexto en el que lo estemos utilizando. Cuando un ordenador tiene que seleccionar una única interpretación de entre varias, para desambiguar se requiere aplicar varias estrategias que sin la ayuda del PLN no se conseguiría [7].

- La **detección de la negación** es un problema de especial relevancia en algunos entornos como, por ejemplo, el jurídico o el médico [8, 9]. Es habitual que se utilicen expresiones negadas con expresiones positivas negadas o al revés, un ejemplo de ello lo podemos ver en la siguiente oración *"no complaints of radicular pain"*. Es un fenómeno lingüístico difícil de detectar automáticamente, pero gracias al PLN se ha conseguido avanzar con buenos resultados en la detección de las negaciones.

- La utilización de **acrónimos** está cada día más extendida. Los acrónimos se definen como las palabras formadas por las iniciales de otras palabras que constituyen la denominación de algo. Su resolución también es uno de los problemas del ámbito del PLN. Esta proliferación de acrónimos hace que los procesos de recuperación y extracción de información puedan verse mermados si no se emplean las técnicas y procedimientos adecuados. En el ámbito de la biomedicina, algunos autores afirman que por cada cinco artículos surge una nueva sigla que llega a coincidir con un gran número de siglas preexistentes y que el uso de los acrónimos hace aumentar la polisemia y la sinonimia léxica, dos fenómenos semánticos que nos llevan de nuevo al problema de la ambigüedad [10].

En la Figura 2.2 se observan las fases más frecuentemente utilizadas en los sistemas PLN [11-13]: i) *análisis morfológico-léxico*, consiste básicamente en la obtención de palabras a partir de un texto, también se realiza la asignación de etiquetas morfológicas a las palabras de un texto (*Part-Of-Speech tagging*); ii) análisis sintáctico: da a conocer las categorías gramaticales de cada palabra y como se combinan los tokens para formar oraciones y textos; iii) *análisis semántico*: esta fase se refiere a la comprensión del lenguaje, se analiza el significado de las palabras y la resolución de ambigüedades léxicas; iv) *análisis pragmático*, se analiza cómo las oraciones se usan en un determinado contexto y cómo afecta al significado de las oraciones.

Figura 2.2. Fases de un sistema basado en PLN

Extracción de Información. El avance y desarrollo de los sistemas PLN [1], junto con la Recuperación de Información (RI) [14], dio origen a otra disciplina dependiente de esta denominada Extracción de Información (EI). La EI, tiene como principal objetivo encontrar y seleccionar información relevante para el estudio de un dominio particular, denominado *dominio de extracción* [15]. Pero quizás una definición más cercana a lo que es hoy en día la EI podría ser la que define esta disciplina como una ciencia que trata de identificar, clasificar y reestructurar información específica existente en fuentes desestructuradas, como por ejemplo los textos, para poder realizar su posterior procesamiento automático [17].

Los primeros estudios relacionadas con la EI se ubican a mediados de los años 60's, pero es a finales de los años 80 cuando esta tecnología comienza a tener auge, lo cual se debe principalmente a tres factores. En primer lugar, el poder computacional que ya empezó a estar disponible con bastante potencia en dicha época; segundo, el exceso de información textual existente en formato electrónico; y por último , la intervención de la Agencia de Defensa de los Estados Unidos (DARPA), que promocionaron durante los años de 1987 a 1998 las siete conferencias de entendimiento de mensajes (MUC) [18] y activaron durante los años de 1990 a 1998 el programa TIPSTER (programa de investigación sobre recuperación y extracción de información del gobierno de EEUU)[19] donde las MUC's fueron incluidas. Las MUC's fueron las que inicialmente fomentaron las competencias entre distintos grupos de investigación. Las cuales se llevaron a cabo con el objetivo de desarrollar sistemas de EI. Muchos sistemas de extracción de información (SEI) surgieron gracias a los MUCs, por ejemplo FASTUS [20], CRYSTAL [21], Autoslog [22], etc. Quizás uno de los sistemas de EI más conocido sea FASTUS (*Finite State Automaton Text Understanding System*), un sistema capaz de extraer información desde texto libre en inglés, japonés y otros lenguajes. Se aplicó inicialmente a la tarea de extraer información de artículos sobre el terrorismo en América Latina para la conferencia MUC-4. Otro foro que ha contribuido históricamente en el ámbito de los sistemas de RI y EI, son las **conferencias TREC** (*Text Retrieval Conference*) [23],

patrocinadas por el NIST (*National Institute of Standards and Technology*) y el Departamento de Defensa de los Estados Unidos, que comenzaron en 1992. En este mismo ámbito de la RI y EI, otro foro que ha aportado gran conocimiento en estas disciplinas son las conferencias ***Cross Language Evaluation Forum* (*CLEF)*** [24]. CLEF es un foro de evaluación que apoya el uso y desarrollo de aplicaciones para la gestión y manejo de librerías digitales. Para ello, desarrollan infraestructuras de prueba, mejora y evaluación de sistemas de recuperación de información multimodal y multilingüe. CLEF nació en enero de 2000, como una evolución de una línea de estudio que se había formado en TREC junto con un grupo de voluntarios europeos, entre 1997 y 1999, para el estudio de los lenguajes multilingües europeos. Todas estas conferencias iniciadas en los años 80 y 90 supusieron un gran avance y un gran marco de referencia para los futuros investigadores en las áreas de la RI y EI.

Algunas de las tareas más importantes llevadas a cabo gracias a los SEI (ver Figura 2.3.) son, entre otras, el reconocimiento de entidades nombradas, resolución de correferencias, reconocimiento de relación entre entidades, reconocimiento de expresiones temporales, etc [25].

Figura 2.3. Arquitectura básica de un SEI

El objetivo final de un SEI es partir de un texto no estructurado y llegar a conseguir, a través de una cascada de módulos que van aportando estructuración al documento, un conjunto de información estructurada y relevante, gracias al filtrado de información a través de la aplicación de determinadas reglas (ver Figura 2.4)[26].

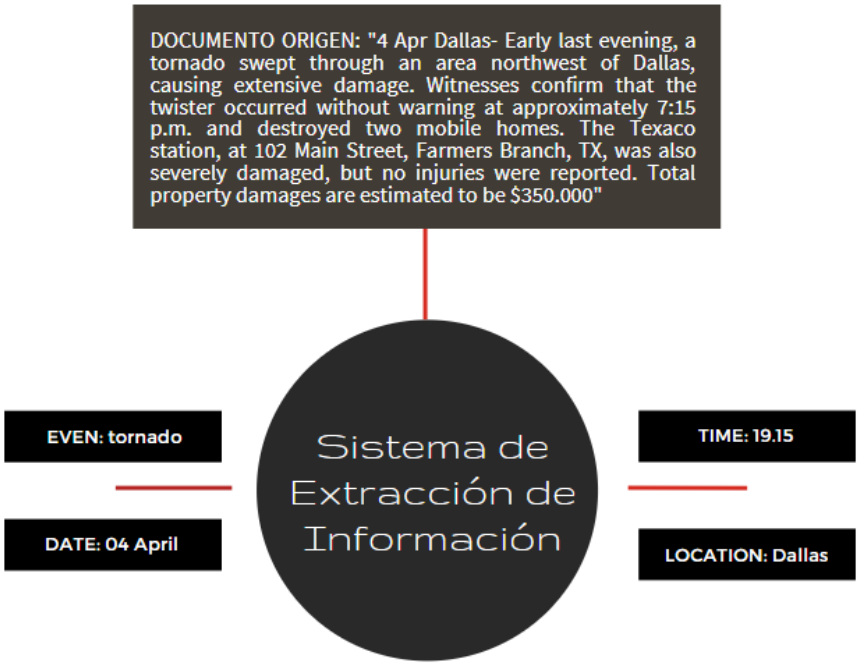


Figura 2.4. Ejemplo de un Sistema de Extracción de Información

Aprendizaje Automático. La disciplina del AA [27], considerada como una rama de la IA, tiene como objetivo la construcción de sistemas capaces de adquirir conocimiento y aprender automáticamente en base a un conjunto de datos de entrenamiento. El AA puede considerarse un proceso de inducción del conocimiento. El auge de la investigación y *workshops* dedicados expresamente a la disciplina del AA tuvieron lugar al inicio de los años 80 (aunque fue en los años 60 cuando surgen los primeros artículos relacionados con AA). Los primeros investigadores en convertir al AA

en una de las subáreas de la IA de mayor importancia y a los que debemos el crecimiento actual de esta disciplina fueron, entre otros, Michalski, Carbonell, Mitchell y Dietterich [27-29].

Una de las primeras definiciones de la palabra “aprendizaje” se recoge en [30] donde se detalla como el aprendizaje marca cambios adaptativos en el sistema que pueden permitir que se realicen las mismas tareas con mayor eficacia cada vez, por tanto, el propósito del aprendizaje es mejorar el rendimiento de algunas clases de tareas [31]. Quizás una de las definiciones más citadas sea la propuesta en [29], donde se afirma que "un programa *aprende de la experiencia* E *respecto a una clase de tareas* T *y una medida de la eficiencia* P*, si su eficiencia en las tareas de* T *se incrementa con la experiencia* E*".* Para Dietterich, el problema de definir aprendizaje se reducía a definir conocimiento [28].

Otros autores más actuales definen el AA como la disciplina de la IA dedicada al diseño de algoritmos para identificar regularidades, patrones o reglas sobre un conjunto de datos o el estudio de algoritmos que pueden aprender relaciones complejas o patrones a partir de datos empíricos y tomar decisiones precisas en base a ellos [32, 33].

Las aplicaciones del AA han sido múltiples [34-36]: análisis de mercado, análisis de riesgos crediticios, detección de fraudes, clasificación de secuencias de ADN, soporte al diagnóstico y pronóstico médico, reconocimiento de patrones, problemas de clasificación, reconocimiento de imágenes, reconocimiento de spam, sistemas de recomendación, soporte a motores de búsqueda, análisis de tendencias, etc. La aplicación de estas técnicas, a lo largo de los últimos 20 años, ha contribuido a mejorar nuestro día a día en prácticamente todos los sectores de la sociedad actual, demostrándose que estas técnicas tienen un alto grado de eficacia y fiabilidad.

En líneas generales, podemos realizar una clasificación de los tipos de AA más comúnmente utilizados según el mecanismo y los métodos que usan para aprender [35,37]:

- Aprendizaje Supervisado
- Aprendizaje No supervisado

El ***aprendizaje supervisado*** [38] es quizás uno de los más empleados en AA, se caracteriza porque el proceso de aprendizaje es controlado gracias a un conjunto de datos de entrenamiento previamente etiquetados, de ahí que reciba el nombre de supervisado. El principal objetivo del aprendizaje supervisado es generar una función capaz de predecir nuevos datos de entrada en base al modelo aprendido que se apoya en un conjunto inicial de datos de entrenamiento. Los pasos necesarios para representar el proceso del aprendizaje supervisado [38] pueden observarse en la Figura 2.5.

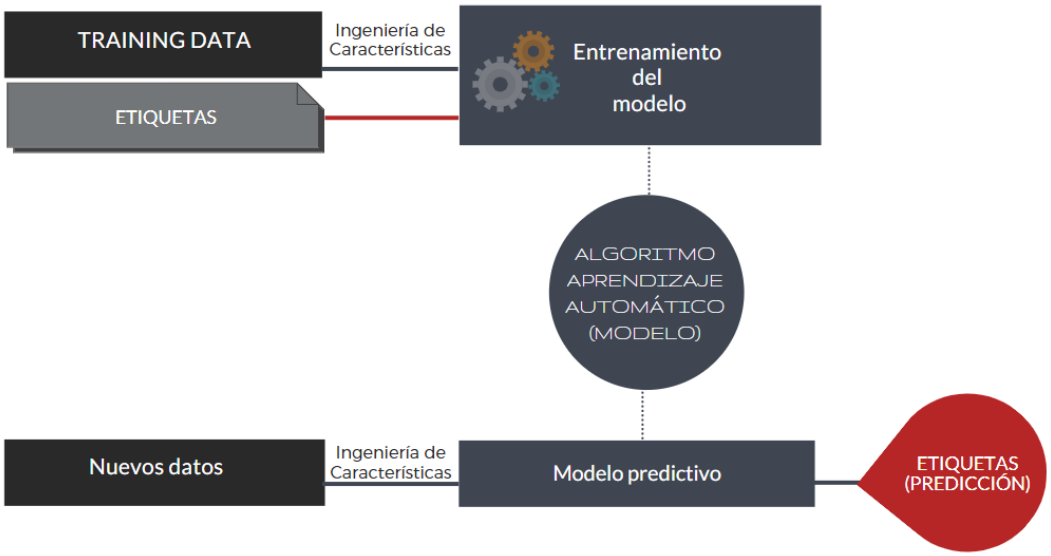


Figura 2.5. Aprendizaje Supervisado

Las modalidades más importantes dentro del aprendizaje supervisado son la clasificación [38] y la regresión [44]. El objetivo de la clasificación es predecir la clase (o clases) a la que pertenece una instancia en base a patrones de entrada previamente etiquetados. Así, a partir de un conjunto entrenado se infiere un modelo (denominado clasificador) que será utilizado para categorizar nuevas instancias no etiquetadas. El objetivo de la regresión es similar al de la clasificación, predecir un valor de salida en base a un patrón de entrada etiquetada, pero se infiere un valor continuo en lugar de categórico. Algunos de los algoritmos de aprendizaje supervisados de uso más extendido son las Máquinas de Vector Soporte (SVM) [68], vecinos más cercanos [73], Naïves Bayes [79], árboles de decisión [83,84] y redes neuronales [91-93] (ver Figura 2.6). Son muchos los trabajos donde se han aplicado las técnicas de aprendizaje supervisado en áreas como la categorización de documentos [70, 80], clasificación de imágenes [71], análisis de sentimientos [40], detección spam [41, 81], detección de enfermedades [97] etc.

En el ***aprendizaje no supervisado*** [45]***,*** al contrario de lo que ocurre con el aprendizaje supervisado, no parte a priori de ningún conjunto de datos etiquetados. El objetivo de este tipo de aprendizaje es realizar agrupaciones de conjuntos de datos similares en base a las características que comparten. Este tipo de aprendizaje es imprescindible cuando se dispone de conjuntos de datos no etiquetados y cuando no se puede asumir el coste necesario para categorizar una gran colección de información.

Una de las tareas más utilizadas dentro del aprendizaje no supervisado es el *clustering*, siendo los algoritmos K-means [69] y DBSCAN [49] los de uso más extendido (ver Figura 2.6). Las técnicas de *clustering* han sido aplicadas en múltiples sectores como segmentación de clientes [46,47], documentación médica [48], apoyo al diagnóstico médico [50], etc.

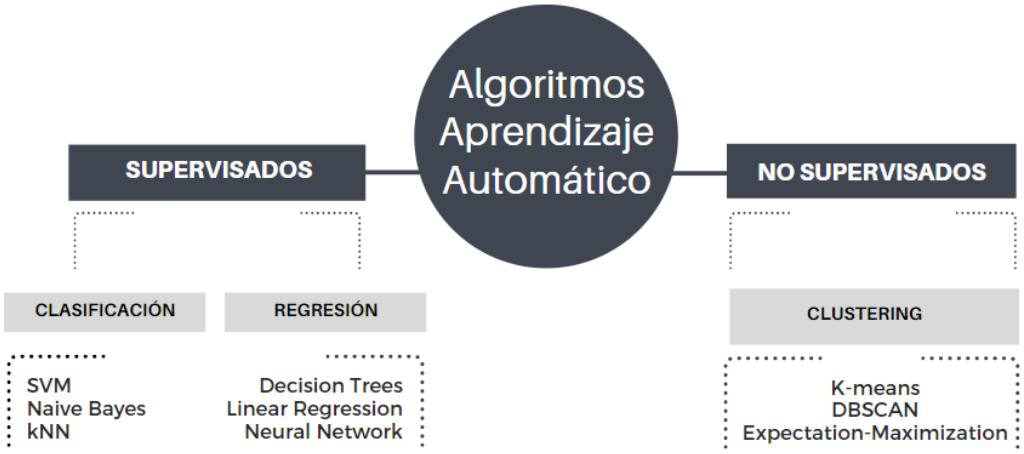


Figura 2.6. Algoritmos Aprendizaje Automático

2.1.3. Fases de un sistema basado en MT

Los sistemas de MT necesitan abordar una serie de fases mínimas para su desarrollo (ver Figura 2.7). Estas fases son ***preprocesamiento, representación y descubrimiento*** [103,104].

Fase de Preprocesamiento. En la fase de ***preprocesamiento*** la colección de documentos se prepara y normaliza para poder ser procesada, puede considerarse como una fase de depuración de la información textual. Es una de las fases más críticas y complicadas ya que en ella se realizan todas las rutinas y métodos necesarios para preparar la información textual para que pueda ser procesada en las siguientes fases de descubrimiento de conocimiento. En esta fase se realizan tareas muy habituales dentro de la disciplina de la MT y del PLN, como pueden ser la tokenización, la lematización, el *stemming* (ver Figura 2.8):

Figura 2.7. Fases de un sistema de MT

a) **tokenización.** El flujo de frases que componen un texto debe ser particionado en componentes más pequeños que tengan un significado. El proceso por el cual se segmentan los documentos procesados en distintas unidades lingüísticas denominados tokens, como frases, palabras, sílabas, etc, es denominado tokenización [105].

b) ***stop-words*** (búsqueda de palabras vacías). Son palabras carentes de sentido por sí mismas, que suelen ser eliminadas en esta fase de preprocesado de los documentos.

Figura 2.8. Técnicas empleadas en la fase de Preprocesamiento

Entran dentro de esta categoría los determinantes, las conjunciones, las preposiciones, etc. [106]. Son palabras comunes que normalmente no contribuyen a la semántica del documento y por tanto, no añaden ningún valor a la interpretación del mismo.

c) **lematización.** Es un proceso de normalización de los términos que componen la colección de documentos. Se basa en la reducción de la palabra a su lema [107,112]. Una de las ventajas que aporta es la reducción de la ambigüedad, aumentando la precisión en la mayoría de los sistemas de MT que lo utilizan [108].

d) ***stemming***. Al igual que la lematización es una técnica de normalización que consiste en reducir una palabra a su raíz [109]. Existen distintos métodos que van desde la simple eliminación del carácter final de cada palabra, hasta otros métodos mejorados y más sofisticados como el conocido algoritmo de Porter [110]. Los métodos de *stemming* son dependientes del idioma, es decir existirá un algoritmo de *stemming* desarrollado para el idioma para el que fue creado. La reducción de la dimensionalidad de los rasgos de las colecciones es una de las ventajas de la aplicación de esta técnica (reducción del tiempo de procesamiento) [204]. Algunos ejemplos del resultado de aplicar la técnica de *stemming* puede observarse en la Tabla 2.1.

Palabras	STEMMING
Persisted Persistence Persisting	Persist
Phenotypic Phenotype	Phenotyp
Lymphocityc Lymphocytes Lymphocyte	Lymphocyt

Tabla 2.1. Ejemplos de aplicación de *stemming*

e**)** ***Part-of-speech-Tagging*** **(POST).** Consiste básicamente en el etiquetado de cada palabra que compone el documento de acuerdo con el papel que juega dentro del texto.

El conjunto más común de etiquetas suelen ser nombre, verbo, adjetivo, artículos, preposiciones, etc [113]. Estas etiquetas proporcionan información semántica de una palabra.

f) identificación de la negación, la detección de conceptos o frases negadas, como por ejemplo "*el paciente no presenta fiebre*", no es una tarea trivial y resulta crucial para la extracción de conocimiento veraz desde la información textual [114, 115].

Fase de Representación. Una vez superada la compleja fase del preprocesamiento de documentos es necesaria la representación del contenido de la colección textual. La representación es una tarea fundamental para el procesamiento automático de documentos. Las representaciones vectoriales son las más sencillas y las más extendidas en MT. En las representaciones vectoriales la idea fundamental es que el significado de un documento se deriva del conjunto de rasgos o características contenidos en el mismo. Dentro de los modelos vectoriales destacan fundamentalmente dos, el índice de latencia semántica y el modelo de espacio vectorial (VSM). El modelo de latencia semántica permite establecer comparativas de similitudes semánticas entre textos. Inicialmente el Análisis de la Semántica Latente fue propuesto por Deerwester, Dumais, Furnas, Landauer y Harshman en 1990, aunque posteriormente fueron Landauer y Dumais cuando en 1997 establecieron una nueva teoría para extraer y representar el conocimiento inmerso en grandes corpus de texto [116, 117]. La teoría del Análisis Semántico Latente (LSA) plantea el uso de un modelo estadístico que permite establecer comparativas de similitudes semánticas entre textos. La co-ocurrencia de palabras es muy importante en este tipo de representación. Este tipo de representación se ha aplicado en diferentes ámbitos y tareas como en la categorización de documentos [118, 119], recuperación de imágenes [120], mejora de buscadores web [121], sumarización de documentos [122, 123], etc.

Pero quizás una de las representaciones textuales más utilizada y básica es la representación vectorial, basada en el VSM, donde cada documento es representado

mediante vectores de términos [124, 125]. En el modelo de representación vectorial los documentos se conforman como un conjunto de rasgos que pueden ser “tratados” y “ponderados”. El modelo vectorial fue propuesto por Salton [126] y permite representar los documentos a partir de un vector de pesos asociados a una serie de rasgos seleccionados del documento. Se asigna un peso a cada término en cada uno de los documentos, en función de la importancia del mismo en cada documento.

Uno de los modelos más utilizados basados en el modelo de espacio vectorial es el llamado ***"bolsa de palabras"*** o "*bag of words*" (BoW) [127]. Para obtener dicha bolsa de palabras de un documento, se convierten en atributos todas las palabras que aparecen en el texto y se les asigna un peso en función de la importancia de cada atributo dentro del documento concreto o de toda la colección de documentos (ver Figura 2.9).
Para asignar este "peso" se necesitan las llamadas *funciones de ponderación* [128,129]. Estas funciones pueden descomponerse en funciones de ponderación local o funciones de ponderación global. Las primeras son aquellas que sólo toman información del documento al que pertenece el rasgo para obtener el peso y las globales son las que toman información de la colección completa para asignar el peso a un rasgo de un documento.

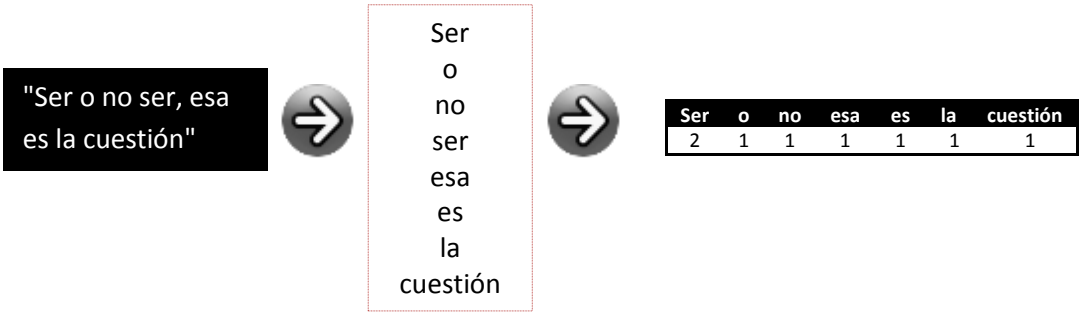


Figura 2.9. Ejemplo Básico "*Bag of Words*"

Algunas de las funciones de ponderación más utilizadas son:

- **Binaria**. Se trata de la forma más sencilla de pesar un rasgo o término y se engloba dentro de las funciones de ponderación local. En este caso el peso del rasgo será 1 si este término aparece en el documento y 0 si no aparece.

$$Bin(t_i, d_j) = \left\{ \begin{array}{l} \text{0, si el término } t_i \text{ no aparece en el documento } d_j \\ \text{1, si el término } t_i \text{ aparece en el documento } d_j \end{array} \right\}$$

- **Term Frequency (TF).** Se trata de otra función de ponderación de tipo local, y determina que un término es más relevante que otro en función de la frecuencia de aparición del rasgo en el documento.

$$\boldsymbol{TF}(t_i, d_j) = f_{ij}$$

siendo f la frecuencia del rasgo t_i dentro del documento d_j.

- **Frecuencia Inversa del documento (BinIDF)**. Se trata de una función de ponderación global, y nos dice que la importancia de un término es inversamente proporcional al número de documentos que lo contienen. Básicamente nos dice que un término que aparece frecuentemente en muchos documentos tiene menor poder de discriminar documentos importantes que otro que aparece con menor frecuencia.

$$\boldsymbol{IDF}(t_i, d_j) = log\ \frac{N}{df(t_i)}$$

siendo N el número de documentos de la colección y $df(\mathrm{t_i})$ el número de documentos donde aparece el término i.

- **TF-IDF (frecuencia del término x Frecuencia Inversa del documento).** Es quizás la función de ponderación global más utilizada en las tareas de MT [190], fue propuesta en 1988 por Salton y Buckey [129]. Este valor combina dos medidas, la frecuencia de la palabra (tf) y la frecuencia inversa del documento (idf). El valor $tf - idf$ se calcula como la frecuencia del término, tf, por la inversa de la frecuencia del documento, idf, que se obtiene dividiendo el número total de documentos del corpus por el número de documentos en los que aparece el término. Los términos con un valor $tf - idf$ más alto son los que mejor caracterizan al documento.

$$\boldsymbol{TF} - \boldsymbol{IDF}\,(t_i, d_j) = f_{ij} \times \log \frac{N}{df(t_i)}$$

Este peso concederá mayor importancia a los rasgos con altas frecuencias en el documento pero con frecuencias bajas en la colección. En el ámbito de la categorización automática de textos, son muchos los estudios que han utilizado la representación vectorial con el peso tf-idf por considerarse una de las funciones que ha ofrecido mejores resultados, además de por su simplicidad y robustez [130].

Así en el estudio realizado por Sahlgren et al. [131] demostraron que en la tarea de clasificación el peso $tf - idf$ mejora los resultados de las funciones *tf* e *idf*. Algunos investigadores han conseguido reinterpretar esta función de ponderación para mejorarla [124, 132]. Una variación de la función *tf-idf* es utilizada actualmente en los motores de búsqueda para identificar los términos claves y poder así filtrar documentos relevantes ante una consulta realizada por un usuario [133, 134].

En conclusión, podemos afirmar que el modelo vectorial, gracias a los buenos resultados obtenidos en numerosas investigaciones relacionadas con el ámbito de la RI, y PLN, junto con su sencillez de aplicabilidad, han hecho que se constituya como uno de los modelos de referencia aplicados en la mayoría de los sistemas de MT. Buena muestra de ello se refleja en los trabajos [135, 136] donde se utiliza el modelo vectorial como base para la construcción de sistemas de MT.

Fase de Descubrimiento. Todas las etapas vistas anteriormente se interconectan para transformar los "almacenes de información" en "fuentes de conocimiento" y, en definitiva, para construir un modelo capaz de descubrir nuevos patrones e información oculta desde grandes volúmenes de contenido textual.

En la última etapa, una vez realizada la etapa de representación, podemos realmente llevar a cabo la fase de descubrimiento, donde se aplicaran distintos algoritmos o métodos (ver Sección 2.1.2), para llevar a cabo las tareas clásicas de la MT [190], como el reconocimiento de entidades nombradas, clasificación y clustering de documentos, resumen automático de textos, creación de ontologías y corpus, etc.

Una vez finalizada la fase de descubrimiento se procederá a evaluar el modelo creado para determinar cuál es su desempeño comparándolo con estándares de referencia.

2.1.4. Tareas de MT: Reconocimiento de entidades nombradas y clasificación automática de documentos.

Dos de las tareas más importantes dentro de la disciplina de la MT, y que han servido de base para el desarrollo de esta investigación, son el reconocimiento de entidades nombradas (NER) y la clasificación automática de documentos (CAD). Analizaremos, a continuación, los puntos claves de cada una de estas tareas.

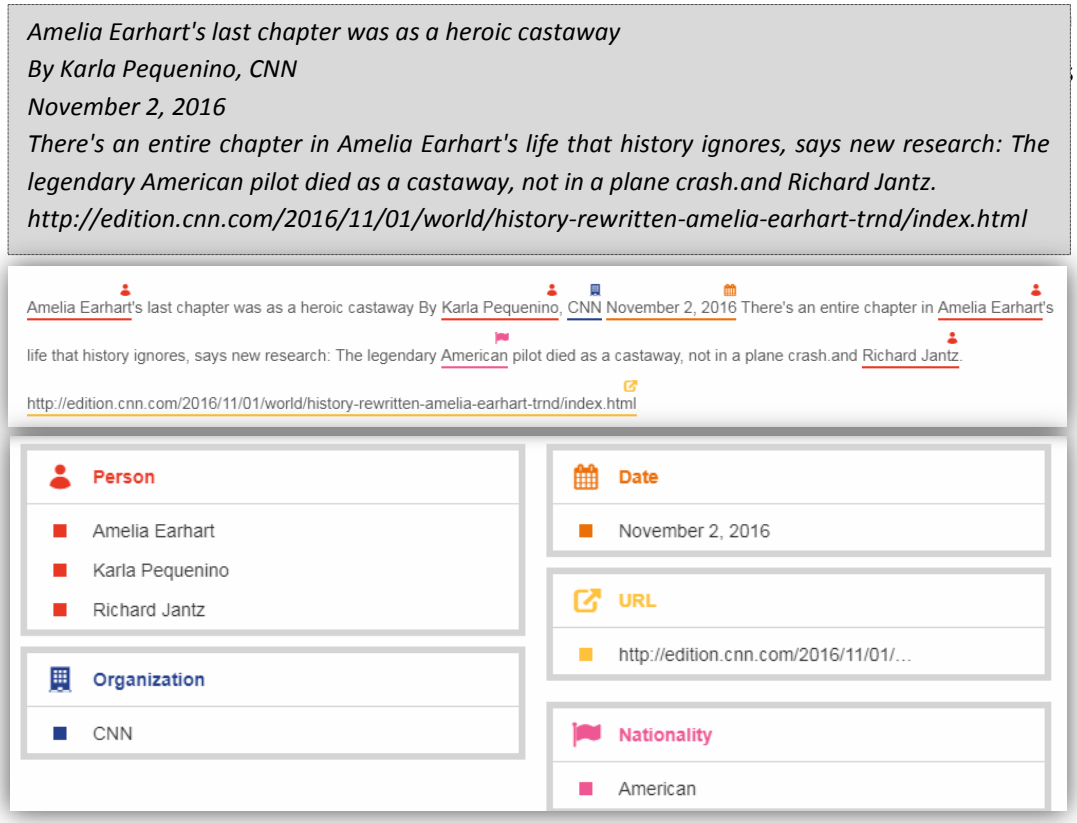
Reconocimiento de Entidades Nombradas. Una entidad nombrada puede definirse como una palabra o conjunto de palabras que se identifican como un nombre de persona, organización, lugar, fecha, tiempo, porcentaje o cantidad. Se llama reconocimiento de entidad nombrada (NER) al descubrimiento de estas entidades nombradas en fragmentos de textos [85, 140]. Por tanto, el objetivo principal de esta tarea consiste en la búsqueda, localización, extracción y clasificación de elementos claves en un texto. Los inicios de los términos "entidades nombradas" y "reconocimiento de entidades nombradas" parten de las conferencias MUC [18]. Estas conferencias se iniciaron a principios de los años 90 para dar cobertura a las investigaciones relacionadas con el área de la extracción y recuperación de información. Fue durante las MUC-6 y MUC-7 donde se centraron en el análisis de las entidades nombradas. Literalmente definieron el concepto entidad nombrada como *"a named object of interest such as a person, organization, or location",* básicamente trataban de buscar y localizar elementos centrales de un texto sobre categoría predefinidas como fueron inicialmente los grupos: personas, organizaciones y lugares.

La finalidad de la extracción de entidades es muy amplia, pasando desde dar soporte a los sistemas de extracción de información, o en la construcción de ontologías, o para dar soporte a los sistemas de sumarización automática, o en apoyo para los sistemas de Búsqueda de Respuestas, etc.

Son muchos los problemas a los que se enfrentan los sistemas NER principalmente por las características especiales del lenguaje natural como son la ambigüedad del lenguaje, la delimitación de la entidad, la detección del idioma, las variaciones lingüísticas, etc. Para entender en la práctica lo que es realmente una entidad nombrada, vamos a partir del siguiente fragmento de texto que será procesado por un sistema NER[3] (ver Figura 2.10). Se obtienen las siguientes entidades nombradas, entre otras, las que identifican personas, fechas, organizaciones, url, etc.

[3] https://www.basistech.com/text-analytics/rosette/entity-extractor/#try-the-demo

La tarea del reconocimiento de entidades nombradas puede ser abordada siguiendo tres enfoques principales: enfoque basado en reglas, enfoque basado en diccionarios y enfoque basado en AA. Los primeros enfoques empleados en los sistemas NER se basaban principalmente en el ***enfoque basado en reglas.*** Estos enfoques utilizan fundamentalmente componentes basados en la aplicación de un conjunto de reglas y heurísticas de extracción que normalmente se centran en características sintácticas o gramaticales (e.g. coocurrencia). Los primeros sistemas basados en reglas surgieron de la conferencia MUC-7, los más destacables son FACILE [174] e IsoQuest [159], dos sistemas que permitían reconocer nombres propios y otras palabras claves encontradas



Amelia Earhart's last chapter was as a heroic castaway
By Karla Pequenino, CNN
November 2, 2016
There's an entire chapter in Amelia Earhart's life that history ignores, says new research: The legendary American pilot died as a castaway, not in a plane crash.and Richard Jantz.
http://edition.cnn.com/2016/11/01/world/history-rewritten-amelia-earhart-trnd/index.html

Figura 2.10.- Ejemplo de reconocimiento de entidades nombradas

en textos y que obtuvieron un alto desempeño (92% para recall-93% precisión para FACILE y 85% en recall-94% precisión para ISoQuest). Algunas investigaciones han demostrado que este tipo de enfoque puede ser robusto y puede funcionar con bastante independencia del dominio, como fue demostrado en la investigación propuesta por Valenzuela-Escárcega et al. [141]. Otros investigadores analizan las ventajas del enfoque basado en reglas para la extracción de entidades nombradas bajo idiomas tan complejos como el árabe [102]. A pesar de estas ventajas, existen detractores de este enfoque ya que aunque se obtienen buenos resultados suelen realizarse bajo entornos restringidos, son muy costosos de implementar debido a que suelen requerir especialistas y un alto tiempo en el desarrollo de las reglas y son muy dependientes del idioma [101].

Los sistemas NER ***basados en diccionarios*** son uno de los más utilizados en la resolución de problemas PLN, estos modelos se basan en contrastar las entidades candidatas con un diccionario que contiene todas las entidades posibles dentro de un ámbito y determinaran si el candidato pertenece o no a una categoría determinada. El uso de este enfoque aporta una gran ventaja ya que en principio ahorra la ardua y laboriosa tarea de anotar corpus y proporciona identificadores únicos para cada candidato (característica muy importante en los sistemas NER que otros enfoques no pueden aportar) [100]. Algunas de las desventajas de este enfoque pueden ser los problemas en la resolución de las variaciones lingüísticas (sinónimos) o la resolución de acrónimos, problemas que pueden hacer que baje la eficacia de estos sistemas NER [95]. Existen un gran número de investigaciones que utilizan el enfoque basado en diccionarios, sobre todo en ámbitos especializados como la Biomedicina o Bioinformática donde es de vital importancia encontrar términos con gran exactitud, donde una coma o un guión desempeña un papel muy importante a la hora de encontrar entidades nombradas [86,94]. En general, es de vital importancia bajo estos enfoques la cobertura de los términos en el diccionario y su especificidad para el ámbito en el que estemos desarrollando el sistema NER.

Los sistemas NER ***basados en AA*** son los más extendidos en la actualidad [85]. El objetivo principal de este enfoque es convertir el problema de la identificación de entidades en un problema de clasificación. Básicamente estos sistemas NER están formados por dos componentes, un conjunto de datos etiquetados para el entrenamiento y un modelo estadístico, de esta forma el sistema identificará y clasificará las entidades en determinadas categorías como nombres propios, localizaciones, fechas, etc en base a este modelo estadístico mediante la utilización de determinados algoritmos de aprendizaje automático.

Existen dos tipos principales de modelos AA que son utilizados en los sistemas NER, métodos supervisados y métodos no supervisados [85]. Los métodos supervisados aprenden un modelo al observar ejemplos anotados en los datasets de entrenamiento, de esta forma, son capaces de predecir el estado de una instancia en base a un conjunto predefinido de instancias previamente etiquetadas. Los principales algoritmos AA utilizados en los sistemas NER bajo el enfoque del aprendizaje supervisado son Conditional Random Fields (CRF) [144], SVM [68] y Hidden Markov Model (HMM) [244]. Los métodos no supervisados [45] son menos utilizados en la construcción de sistemas NER, estos modelos aprenden sin necesidad de partir de colecciones de datos previamente categorizadas. Uno de los enfoques más típico dentro de este tipo de modelos es el clustering (agrupación de entidades nombradas en base a similitudes en el contexto).

Los principales problemas que generan los métodos basados en AA se basan en que se requieren expertos y tiempo para anotar las colecciones de partida necesarias para entrenar los modelos supervisados. Aunque este problema se puede paliar con la utilización de métodos no supervisados, pueden seguir sin obtener un rendimiento adecuado en entornos con un vocabulario tan complejo y especializado como la Medicina.

Clasificación Automática de Documentos. La *CAD* puede definirse como un proceso de aprendizaje por el cual se pueden inferir determinadas características que diferencian un documento para poder así distinguir a que clase o categoría pertenece [225,227]. La tarea CAD, una de las tareas de MT más ampliamente utilizada, tiene como objetivo categorizar automáticamente una colección de documentos (con una o más clases), en base a un conjunto de datos de entrenamiento con clases previamente conocidas. Dado un conjunto de datos $D = \{d1, d2, \dots, dn\}$ formado por un conjunto de documentos y un conjunto de clases $C = \{C1, C2, \dots, Cm\}$, el problema de la clasificación es encontrar una función $f: D \rightarrow C$, tal que cada d_i sea asignado a una clase C_j [227]. El enfoque básico de la CAD se representa en la Figura 2.11, donde partiendo de un conjunto de documentos previamente categorizados se construirá un modelo que servirá de base para clasificar nuevos casos que previamente se encontraban sin clasificar [228].

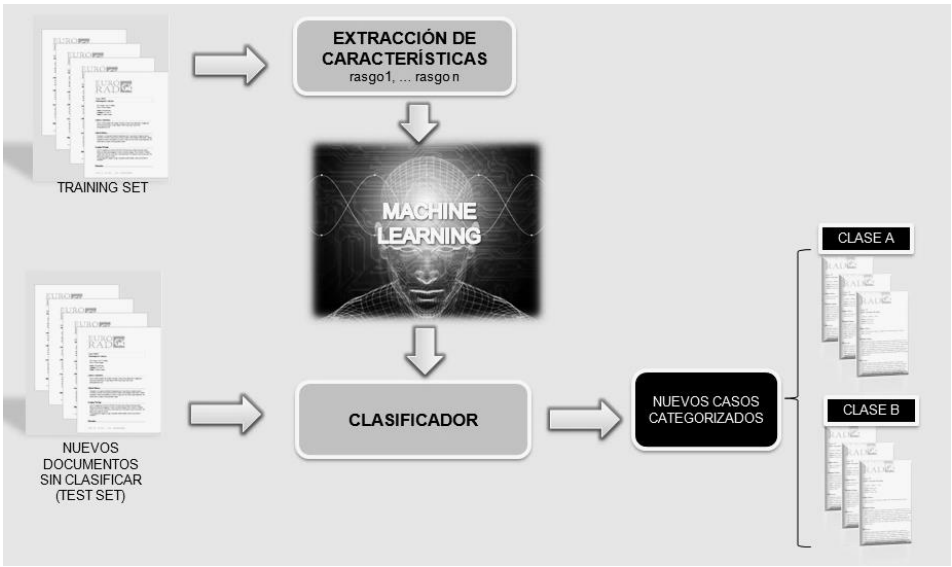


Figura 2.11. Proceso básico de clasificación automática de documentos basada en ML

En virtud al número de clases, la clasificación de documentos puede abordarse desde las siguientes perspectivas: binaria, multiclase y multietiqueta. La clasificación ***binaria*** es una de los modelos de clasificación más simples. A través de este tipo de clasificación vamos a poder determinar la pertenencia o no pertenencia a una clase (verdadero/falso, si/no), cada instancia puede recibir una de dos posibles etiquetas y en ningún caso puede existir solapamiento de clases. La clasificación ***multiclase*** se da cuando una instancia puede ser asignada a una categoría dentro de un conjunto múltiple de clases (mayor a dos). Pero la mayoría de los problemas reales de clasificación a los que nos enfrentamos en la vida real no pueden abordarse desde una perspectiva tradicional, ya que existen un gran número de colecciones de datos no convencionales, donde una instancia (documento) puede pertenecer a más de una categoría al mismo tiempo. A este tipo de clasificación se le denomina **clasificación *multietiqueta*** (MLC, por sus siglas en inglés) [229]. La MLC es un paradigma de aprendizaje por el cual una instancia (documento) puede ser asignada a más de una clase o categoría, donde los conjuntos de etiquetas no son excluyentes entre sí (ver Figura 2.12). Es una de las principales áreas de interés en el ámbito de la MT y AA, y es una de las tareas centrales de esta investigación.

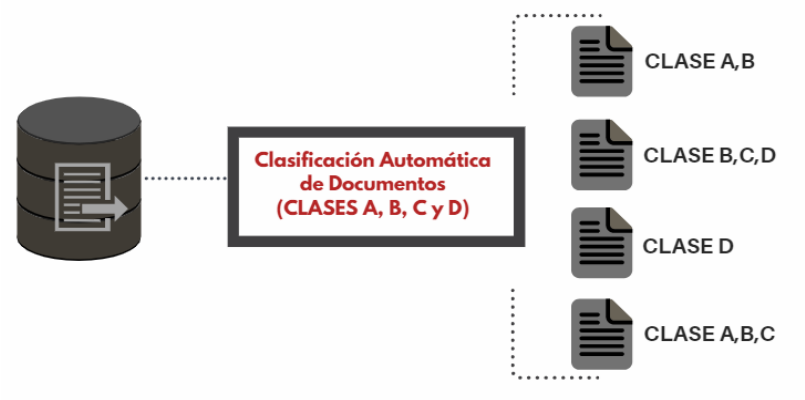


Figura 2.12. Clasificación Multietiqueta

Estos problemas complejos, conocidos como problemas de clasificación multietiqueta, se abordan mediante la aplicación de diferentes métodos que pueden agruparse en tres grandes grupos: ***métodos de transformación de problemas, métodos de adaptación de algoritmos y métodos basados en multiclasificadores*** [231]. En la Figura 2.13 se recoge un esquema de los principales métodos MLC.

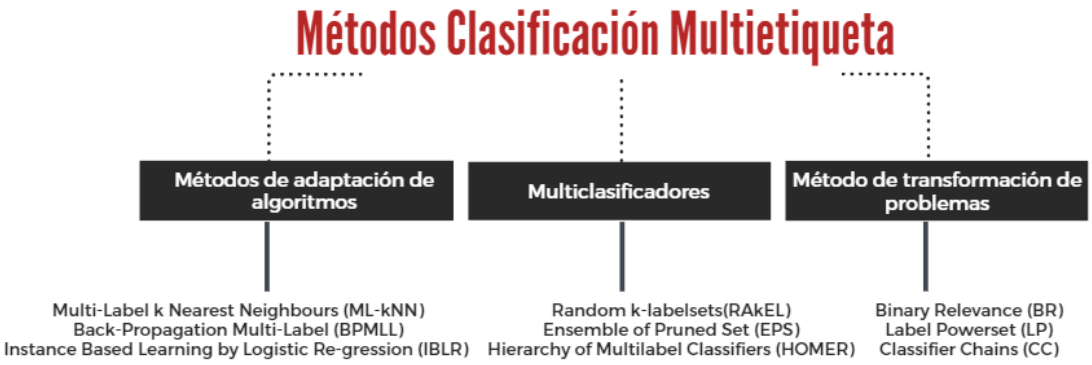


Figura 2.13. Principales Métodos Clasificación Multietiqueta

Los métodos de transformación de problemas tienen como objetivo convertir un conjunto de datos multietiqueta en uno o varios conjuntos de datos de una única etiqueta. Algunos de los enfoques más utilizados en la literatura dentro de los métodos de transformación de problemas son Binary Relevance (BR), Label PowerSet (LP) y Classifier Chains (CC). ***El método BR*** es el método de transformación de problemas más popular, cuyo propósito es transformar los problemas de clasificación multietiqueta en problemas de clasificación binaria, generando un conjunto de datos binarios para cada etiqueta. A pesar de algunas desventajas del método (considera que las etiquetas son independientes entre sí [232]), han demostrado que sigue siendo eficiente y competitivo frente a otros métodos más complejos cuando aumenta la complejidad del conjunto de datos [58]. El ***método LP*** crea una nueva clase por cada posible combinación de

etiquetas. Al contrario de lo que ocurre con BR, este método si tiene en cuenta la relación entre etiquetas [233, 234]. El ***método CC,*** está basado en el método BR pero supera las desventajas del mismo consiguiendo aumentar el rendimiento predictivo aunque mantiene un elevado tiempo de computación. Este método genera q clasificadores binarios encadenados para que cada clasificador incluya como entradas las etiquetas predichas por los clasificadores anteriores.

Los métodos de adaptación de algoritmos emplean técnicas de clasificación tradicionales que han sido adaptadas para trabajar con datos multietiqueta. Algunos de los más significativos de este grupo son AdaBoost, Multi-Label K-Nearest Neighbors (ML-kNN) y instance-based learning by logistic regression (IBLR). ***AdaBoost*** es el algoritmo boosting más popular y altamente competitivo en la categorización de textos, una de sus extensiones más conocidas es AdaBoost.MH. Este método realiza una reducción del problema multietiqueta mapeando cada ejemplo a q ejemplos binarios, uno para cada etiqueta del conjunto de datos, posteriormente AdaBoost se aplicará a estos datos binarios. ***ML-kNN*** es una adaptación del popular algoritmo de aprendizaje K Nearest Neighbors (kNN) para datos multietiquetas. Dada una instancia de prueba, este método determina en primer lugar sus k vecinos más cercanos en el conjunto de entrenamiento. A continuación, utiliza el principio de máximo a posteriori para determinar el conjunto de etiquetas de la instancia de prueba de acuerdo con la información estadística obtenida de los conjuntos de etiquetas de las instancias vecinas. Este método tiene en cuenta la correlación de etiquetas. ***IBLR*** es una combinación del aprendizaje basado en instancias y las técnicas de regresión logística. Este método considera las etiquetas de las instancias vecinas como características adicionales de dicha instancia, reduciendo por lo tanto el aprendizaje basado en instancias a un problema de regresión logística.

Los métodos basados en multiclasificadores, también denominados métodos ensembles, combinan la respuesta de varios clasificadores. Entre los más empleados se encuentran RAmdom k labELsets (RAkEL), Ensemble of Pruned Set (EPS), Ensemble

Classifier Chain (ECC), multi-label stacking (MLS) y hierarchy of multi-label classifiers (HOMER). ***RakEL*** combina varios clasificadores *LP.* Bajo este método, cada clasificador es entrenado con un subconjunto aleatorio (sin reemplazamiento) de k etiquetas empleando LP. Para una nueva instancia, la respuesta de los clasificadores se promedia por etiqueta y por último, se considera un proceso de votación para designar las etiquetas asignadas. ***EPS,*** basándose en el método Pruned Sets, crea varios clasificadores de conjuntos podados donde cada clasificador se entrena con una muestra del conjunto de entrenamiento sin reemplazo. El método ***ECC*** entrena un conjunto de clasificadores CC, cada uno con un orden de cadena aleatorio y un subconjunto aleatorio de patrones con reemplazo del conjunto entrenamiento. ***MLS***, también conocido como 2BR, ya que aplica el método BR dos veces, primero entrena un clasificador BR y a continuación, utiliza las salidas de los clasificadores anteriores como entradas para entrenar un nuevo clasificador. Por último, el método ***HOMER*** construye una jerarquía de clasificadores multietiqueta, donde el conjunto de entrenamiento inicial se divide en varios grupos (aplicando una metodología de clustering), dando lugar a un clasificador basado en árboles con un número cada vez menor de etiquetas.

2.2. Minería de Textos en el dominio de la Medicina

Actualmente uno de los grandes desafíos computacionales en el ámbito de la Medicina es poder obtener el conocimiento implícito inmerso en la gran maraña de información clínica textual que se genera diariamente en los centros sanitarios. Como hemos visto en la introducción, esta sobreabundancia de información no está aportando actualmente ese "plus" de apoyo que necesitan los profesionales sanitarios en su labor clínica diaria para facilitar la toma de decisiones.

Hasta la fecha, en Sanidad se ha invertido mucho en construir Sistemas de Información que recopilan y almacenan toda la información asistencial generada en los centros sanitarios, ahora ha llegado el momento de invertir en extraer Inteligencia y valor de

todos estos depósitos textuales y convertir estos "almacenes de información" en "fuentes de conocimiento" (Figura 2.14). Se necesitan herramientas ágiles y eficaces que sean capaces de extraer conocimiento de estos contenedores de información textual para conseguir facilitar la labor de los sanitarios en tareas como, la identificación automática de factores de riesgo para prevenir la aparición de enfermedades, para apoyar la resolución y búsqueda del mejor diagnóstico, para identificar conflictos entre distintos tratamientos del paciente, para evitar posibles errores en las interacciones medicamentosas, para realizar estudios predictivos personalizados, etc.

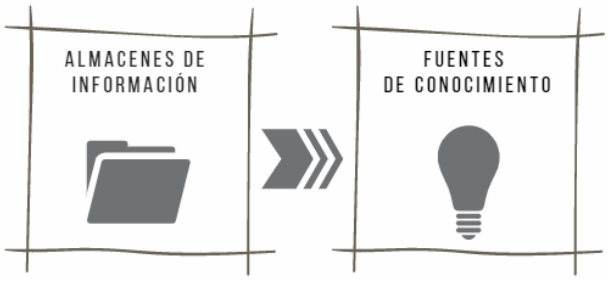


Figura 2.14. Almacenes a Conocimiento

En la actualidad para abordar la problemática de extraer valor del texto clínico, en el entorno de la medicina computacional, disponemos de las técnicas avanzadas que nos proporciona la disciplina de la MT. A continuación, analizaremos con profundidad las principales aportaciones de la MT a la Medicina, analizando un gran número de estudios relevantes sobre la aplicación de sus técnicas en el ámbito asistencial, profundizaremos en las dos tareas centrales de nuestra investigación como son el reconocimiento de entidades médicas y clasificación diagnóstica y por último, analizaremos los principales recursos y herramientas que facilitan el análisis de colecciones textuales en el ámbito de la Medicina.

2.2.1. Evolución de las tareas de MT en Medicina

En las Jornadas Big Data y Analytics en el Sector Sanitario[4], algunos expertos del área de la Salud plantearon varias cuestiones en relación a la información sanitaria,

- "el 80% de los datos sanitarios contiene información no estructurada"
- "sólo el 10% de las organizaciones sanitarias usan herramientas de análisis avanzados para extraer información"
- "la situación de tensión por la que atraviesa el sistema sanitario debido a la falta de personal sanitario junto con el envejecimiento de la población puede destensionarse gracias al apoyo de la revolución tecnológica"

Son muchos los profesionales de la sanidad que ya reclaman una gestión inteligente de la información para poder obtener valor de la principal fuente de información que es la historia clínica electrónica. Aunque es evidente la complejidad de este reto, extraer valor del texto clínico, la disciplina de la MT, junto con otras disciplinas imprescindibles como el PLN y el AA, han permitido afrontarlo con excelentes y prometedores resultados, en la última década.

La MT se ha convertido en la actualidad en una herramienta muy confiable y eficaz en el ámbito médico [138]. Desde principios de siglo, cuando aparecieron los primeros trabajos de investigación, las técnicas de MT aplicadas a tareas relacionadas con el ámbito de la Medicina (diagnóstico, tratamiento y prevención de enfermedades) están denotando un interés progresivo [137], como lo demuestra el número creciente de artículos publicados al respecto en las bases de datos biomédicas Embase y PubMed (ver Figura 2.15). El número de competiciones relacionadas con la biomedicina y la MT han aumentado en los últimos años (en promedio, más de cuatro desafíos diferentes por año a partir de 2008), lo que produjo un incremento de artículos de investigación al respecto [138].

[4] http://www.pmfarma.es/noticias/22420-big-data-y-analyitics-en-sector-sanitario-para-mejorar-la-vida-de-pacientes.html

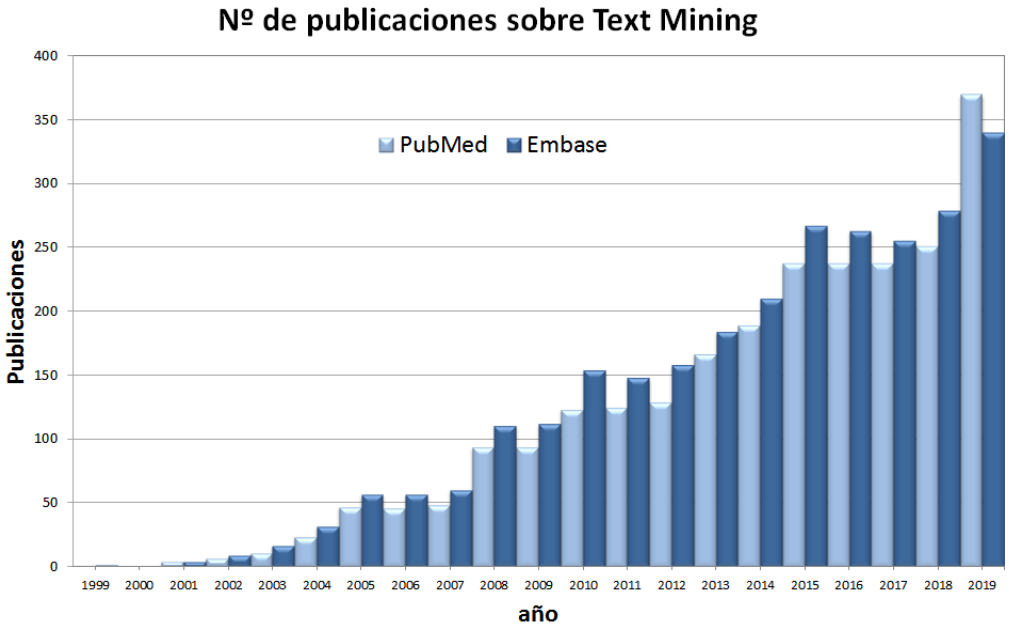


Figura 2.15. Nº Publicaciones según bases de datos biomédicas PubMed y Embase (palabras claves: Text Mining). Última actualización Febrero 2020.

La MT es una de las disciplinas que, junto con el PLN y el AA, más ampliamente han sido utilizadas en los últimos diez años en el área de la Biomedicina [138]. Las investigaciones relacionadas con este campo han estado orientadas a tareas como la extracción y recuperación de entidades biológicas (genes) desde la literatura científica, estudio de las interacciones de estas entidades (gen-enfermedad, gen-proteína, etc), análisis de microarrays, extracción de terminología biomédica, farmacogenómica, etc [160]. Pero tras la explosión de las investigaciones biomédicas, empiezan a cobrar protagonismo las aplicaciones de la MT al área de la Medicina Clínica, orientada al diagnóstico, tratamiento y prevención de la enfermedad.

Un amplio número de investigaciones se han centrado, en los últimos años, en las tareas de MT que analizaremos a continuación:

Descubrimiento de conocimiento y generación de hipótesis. La identificación de eventos clínicos relevantes, ocultos en grandes volúmenes de datos textuales, y la generación de hipótesis es una de las aportaciones más importantes de la MT a la Medicina. El descubrimiento de este conocimiento oculto inmerso en textos clínicos no estructurados va a permitir al profesional sanitario concentrarse en la información más relevante y crítica del paciente, consiguiendo así apoyar al clínico en una toma de decisiones más precisa en un menor tiempo, un factor muy determinante en Medicina. Una de los primeros investigadores que pusieron en práctica la base de la MT y el descubrimiento de nuevas hipótesis médicas fue *Swanson* [175, 176]. Swanson y Smalheiser, partiendo de la información textual contenida en la literatura médica (artículos científicos) y utilizando técnicas de la MT, consiguieron descubrir nuevas hipótesis que generaron un conocimiento muy importante, *la relación entre la migraña y el magnesio.* Las hipótesis principales extraídas de este estudio fueron: i) El estrés está asociado con la migraña; ii) el estrés genera pérdidas de magnesio; iii) Los pacientes con migraña tienen una alta agregación de plaquetas. Todas estas hipótesis llevaron a la conclusión de que existía un vínculo entre la falta de magnesio y algunos tipos de migraña, un hallazgo hasta la fecha sin descubrir.

El descubrimiento y la identificación temprana de los factores de riesgo de algunas enfermedades graves, como el cáncer o las enfermedades cardiacas, son ejemplos de la importancia de la aplicación de las técnicas de la MT en el ámbito de la Medicina. Una prueba de ello lo podemos encontrar en el trabajo realizado por *Byrd et al*. donde se presenta un modelo capaz de descubrir los signos y síntomas tempranos de pacientes que pueden desarrollar una insuficiencia cardiaca [177]. Se utilizaron dos enfoques, uno basado en AA y otro basado en reglas lingüísticas. Los autores demostraron que la tarea de extracción de criterios de Framingham desde textos clínicos alcanzaron unos valores altos en precisión y recall, lo cual indica que puede ser un buen método de apoyo al clínico para mejorar la detección temprana de la insuficiencia cardíaca.

Collier presentó un trabajo cuyo objetivo era ofrecer una visión genérica del papel que han desempeñado las técnicas de la MT en el área de la Salud Pública, y en concreto, en la detección de epidemias [179]. El autor destacó la contribución de la MT en la detección de alertas de riesgos para la salud pública y nos mostró como la utilización de estas técnicas hacen posible el desarrollo de herramientas de gran utilidad, como Biocaster. Biocaster es un interesante servicio Web que se basa en la aplicación de técnicas del PLN y la MT para ofrecer la detección y el seguimiento de brotes de enfermedades infecciosas a través de un mapa mundial visual [180].

Los autores *Baron et al.* realizaron un meta-análisis para identificar los efectos adversos del uso de la aspirina [181]. Para ello, partieron de un total de 119,310 publicaciones para identificar ensayos clínicos o estudios observacionales donde se compararon la toxicidad gastrointestinal de la aspirina con otros medicamentos (incluidos un placebo). Ante la imposibilidad de realizar este descubrimiento de información en millones de artículos y documentos textuales de forma manual, se emplearon técnicas de MT para seleccionar artículos relevantes para este meta-análisis.

Después de esta búsqueda selectiva, fueron seleccionados 150 ensayos clínicos relevantes, de los que se concluyeron que los eventos adversos gastrointestinales severos eran muy raros y escasos por el uso de la aspirina, aunque se demostró que la aspirina confiere un mayor riesgo (50-100%) en problemas gastrointestinales menores como nauseas, vómitos, dolor abdominal o dispepsias. Se demuestra una vez más como el descubrimiento de eventos clínicos obtenidos desde información textual implícita y oculta en grandes colecciones de datos, puede ayudar en la construcción de nuevas hipótesis médicas.

La evaluación y gestión del dolor en pacientes con cáncer de próstata metastásico es analizado por *Heintzelman et al*. [184]. Se presentó un sistema automatizado, capaz de clasificar y hacer un seguimiento del dolor del paciente a través de los registros médicos electrónicos, que ha permitido mejorar la atención clínica y poder identificar nuevos fenotipos. Para la realización de este sistema se utilizaron técnicas de MT, metatesauros

y PLN. En esta investigación, los hallazgos encontrados en la cohorte inicial sugieren que pueden existir fenotipos "*outlier*" del dolor, muy útiles para la investigación sobre la base molecular del dolor causado por cáncer.

Otras investigaciones relacionadas con el descubrimiento y generación de hipótesis médicas donde se han aplicado la MT pueden encontrarse en [182, 183, 185-188].

Generación automática de resúmenes o sumarización. La generación automática de resúmenes [190, 191] ha sido hasta hoy ampliamente utilizada en la Biomedicina y actualmente está creciendo su utilización en la disciplina de la Medicina Clínica. La dificultad a la que se enfrentan la mayoría de facultativos para estar al día y actualizados en los temas más novedosos en Medicina es cada vez mayor. Miles de nuevos ensayos, artículos, guías clínicas, revistas, noticias, etc se publican diariamente, la lectura de toda esta información es prácticamente inabordable si no se utilizan técnicas como la sumarización. En *Elhadad et al*. analizaron la problemática de la sobreabundancia de información médica y como la generación automática de resúmenes puede ayudar a paliar este problema [192]. Se presenta un sistema llamado PERSIVAL (*PErsonalized Retrieval and Summarization of Images, Video and Language*) que tiene como objetivo proporcionar al facultativo presentaciones a medida de la literatura médica más relevante. Los autores *Fiszman et al.* recogen el desarrollo de un sistema de resumen automático para ayudar a los facultativos a encontrar información relevante correspondiente a algunas enfermedades en base a los resultados de búsquedas en PubMed [193]. Otros autores como *Workman et al.* evalúan y analizan la importancia de Semantic MEDLINE [200] como una herramienta de apoyo en la toma de decisiones médicas y aportan nuevas investigaciones para mejorar esta herramienta [201]. Otras investigaciones que abordan el problema de la sumarización de documentos clínicos se recogen en [198, 199, 203, 205].

Extracción terminológica para la construcción de ontologías, corpus y bases de datos. Las ontologías, tesauros y corpus son recursos muy importantes dentro del contexto sanitario para complementar tareas como la recuperación y extracción de información, búsqueda de respuesta, clasificación automática de documentos, etc. En *Roberts et al.* describen la construcción de un corpus (CLEF corpus[5] [208]) anotado semánticamente desarrollado en base a textos clínicos, cuyo propósito es apoyar la evaluación de sistemas de extracción automática de información clínica [207]. Para la construcción del sistema se utilizó GATE NLP toolkit [209]. En el trabajo realizado *por Islamaj et al.* presentaron un corpus creado con la idea de ayudar al desarrollo y evaluación de la tarea de reconocimiento de nombres de enfermedades [210]. El corpus NCBI fue anotado manualmente en base a una colección de 793 abstracts de PubMed, contiene 6.892 menciones de enfermedades y 790 conceptos de enfermedad únicos. En esta investigación los autores emplearon la herramienta MetaMap [155] para identificar los conceptos UMLS [195] encontrados en los textos de PubMed. En el trabajo realizado *por Fabian et al.* desarrollaron un enfoque para la ampliación de ontologías [211]. Se emplearon dos métodos distintos: un enfoque basado en texto, utilizando técnicas de la MT y otro basado en la estructura HTML de los documentos. El enfoque fundamentado en técnicas de MT se basó en la extracción de patrones del texto para la obtención de nuevos candidatos a términos. En *Fang et al.* analizan la construcción de una base de datos denominada TCMGeneDIT[6], que proporciona información sobre las relaciones entre la Medicina China Tradicional, los genes, las enfermedades y los efectos de la medicina tradicional que han podido obtenerse desde la literatura científica, basándose en técnicas de la MT [222]. Otras investigaciones relacionadas, basadas en MT, se presentan en [212, 215-217].

[5] http://www.clef-initiative.eu/dataset/corpus
[6] http://tcm.lifescience.ntu.edu.tw/index.html

Además de las tareas detalladas anteriormente, existen dos tareas de la MT fundamentales en esta investigación e imprescindibles para el apoyo a la toma de decisiones clínicas, estas son el reconocimiento de entidades médicas y la clasificación diagnóstica. Debido a su importancia, profundizaremos detenidamente en ellas en las siguientes secciones.

2.2.2. Reconocimiento de Entidades Médicas

Dentro del ámbito de la Medicina se suele denominar a los sistemas NER (ver Sección 2.1.4) como sistemas MER (*Medical Entity Recognition*) [142]. El reconocimiento de entidades nombradas en Medicina se aplica a la detección y extracción de conceptos de interés, como el nombre de una enfermedad, un síntoma o una región anatómica, extraídos de una colección de informes clínicos o de un conjunto de episodios clínicos informatizados (ver Figura 2.16). Una de las principales ventajas de esta tarea, es aportar apoyo al clínico en la extracción de conocimiento relevante que puede pasar desapercibido entre la gran cantidad y variabilidad de fuentes de información médicas textuales que el profesional sanitario maneja en su actividad diaria. Estos sistemas han

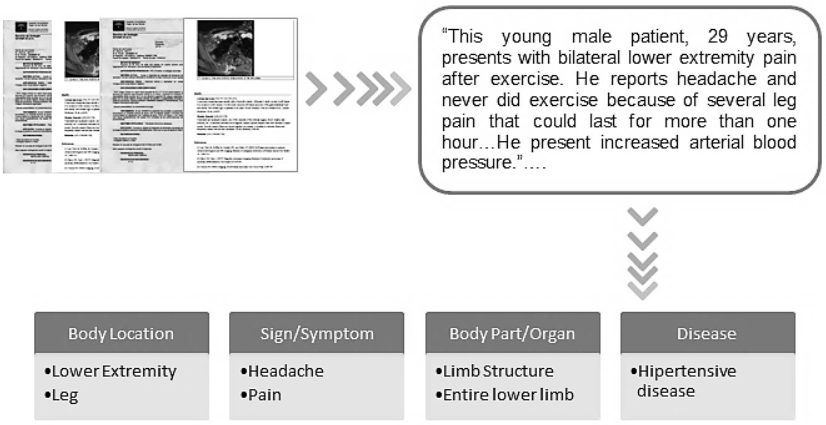


Figura 2.16. Extracción de Entidades Médicas desde Informes radiológicos

realizado una importante contribución en el ámbito de la Medicina, ya que han permitido aportar información de gran valor para el clínico como la detección de alertas clínicas localizando alergias, reacciones adversas a medicamentos o problemas médicos, basándose exclusivamente en información textual [178]. Tal y como analizamos en la Sección 2.1.4, los sistemas MER pueden abordar la tarea del reconocimiento de las entidades médicas desde tres enfoques [140]: enfoque basado en reglas, enfoque basado en diccionarios y enfoque basado en ML.

Para demostrar la importancia de los sistemas MER en el ámbito de la Medicina, se analizaran un amplio número de investigaciones que se centran en la extracción de conceptos médicos desde informes clínicos textuales, el descubrimiento de conceptos relacionadas con la expresiones temporales, anonimización de datos personales y por último, se recogerán la principales investigaciones en relación a la extracción de relaciones entre entidades médicas.

Extracción de Entidades Médicas desde informes clínicos. En el trabajo presentado por *Roberts et al.* se analizan dos técnicas: identificación de conceptos médicos en un texto clínico y la clasificación de afirmaciones que pueden indicar la existencia, ausencia o incertidumbre de un problema médico [143]. Se utilizó un sistema NER basado en AA para la extracción de clases semánticas como, nombre de enfermedades, medicamentos y proteínas, extraídos de los informes de alta y notas de evolución. Los autores emplean dos clasificadores para llevar a cabo el experimento: SVM [68] y CRF [144].

En el trabajo que presenta *Aparicio et al.* se presenta una herramienta capaz de extraer conceptos médicos relevantes desde textos clínicos haciendo uso del reconocimiento de entidades nombradas [145]. Este sistema, que sigue un enfoque basado en diccionarios y otras fuentes de conocimiento como las ontologías, consigue optimizar el proceso de

localización de información médica sobre casos clínicos para distintos tipos de perfiles de usuarios (médicos, investigadores, estudiantes, etc).

La clínica Mayo en un estudio recogido en [146], propone la evaluación de un sistema NER, basado en diccionarios, que forma parte de su sistema de extracción de información. En concreto, analizan las entidades enfermedad y trastornos, basándose en sus resultados demostraron la complejidad de la identificación de entidades médicas debido fundamentalmente a fenómenos lingüísticos como la sinonimia o la negación. Actualmente la Clínica Mayo está investigando con la incorporación de un enfoque basado en AA, con la utilización de SVM y CRF para aplicarlos al reconocimiento de entidades médicas.

El objetivo de la investigación realizada por *Jiang et al.* es desarrollar un sistema NER para extraer problemas médicos, pruebas y tratamientos desde los informes de alta hospitalaria escritos en lenguaje natural [147]. Esta investigación dio como resultado la creación de un sistema híbrido (AA/Reglas NLP) con unos resultados prometedores en la tarea del reconocimiento de entidades.

Xu et al. presentan la creación de un sistema capaz de extraer automáticamente entidades nombradas desde informes radiológicos [148]. Para la construcción de este sistema combinaron un clasificador LSP (*labeled sequential pattern*) [149] y una técnica ampliamente utilizada para el reconocimiento de entidades llamada CRF [144]. Este trabajo demostró que es posible utilizar esta metodología para generar sistemas de alertas automáticas mediante la detección de determinados fragmentos textuales de los informes radiológicos. La incorporación de fuentes externas de conocimiento a los sistemas NER ha aportado una mejora en los resultados obtenidos para la tarea de detección de entidades dentro del dominio de la Medicina.

En el trabajo realizado por *Bodnari et al.* podemos comprobar como la combinación del modelo CRF junto con la utilización de un conjunto de características alimentado y expandido por fuentes externas como Wikipedia, MetaMap [155], UMLS [195] y cTakes [151], pueden mejorar los valores de precisión [150]. Otro estudio que avala la

mejora de la tarea de detección de entidades gracias a la combinación de varias fuentes externas de conocimiento, puede verse en el trabajo realizado por *Xia et al.* [152]. En esta investigación se combinan las herramientas MetaMap y cTakes en la tarea del reconocimiento de enfermedades en textos clínicos, tarea propuesta por la conferencia CLEF eHealth 2013. Los autores generaron dos sistemas baseline, uno basado en la extracción de características del sistema MetaMap y otro basado en la extracción de características del sistema cTakes, ambas herramientas incorporan la utilización de metathesaurus. Para evaluar ambos sistemas construyeron un sistema combinado, ya que el sistema baseline MetaMap obtuvo mejores resultados en precisión y el sistema baseline cTakes obtuvo mejores resultados en recall. Los resultados de la evaluación mostraron que el sistema combinado puede funcionar mejor que los sistemas baseline funcionando individualmente.

La amplia mayoría de los sistemas MER desarrollados en la actualidad tiene un buen rendimiento cuando funcionan sobre textos clínicos en inglés, pero ¿qué ocurre cuando este contenido textual se encuentra escrito en otro idioma? ¿son tan efectivos estos métodos?. Estas preguntas fueron resueltas en el estudio desarrollado por *Skeppstedt et al.* donde intentaron demostrar la eficacia de la extracción de entidades relevantes de textos clínicos escritos en sueco (enfermedad, hallazgos, farmacología y órgano) [153]. Como algoritmo de AA fue seleccionado el CRF por haber sido uno de los más empleados en la tarea de extracción de entidades desde informes clínicos. Este estudio demostró que los métodos MER aplicados con anterioridad a textos en inglés también obtuvieron resultados satisfactorios sobre textos clínicos en sueco.

La extracción de entidades nombradas en textos clínicos escritos en español es una tarea compleja y a su vez poco investigada. En el trabajo desarrollado por *Carrero et al.,* [154] se propuso la creación de un sistema basado en la herramienta MetaMap Transfer (MMTx) [155] para la extracción de conceptos desde textos médicos en español. Actualmente esta herramienta no permite la explotación de textos en español. Para solventar este problema propusieron combinar las técnicas de traducción automática

junto con el uso de las ontologías biomédicas para llegar a producir un texto en inglés que pudiera ser procesado por MMTx. Los resultados demostraron que el uso de traductores automáticos produjeron una alta similitud con los textos originales, lo cual abre una vía importante para la realización de la tarea de extracción de conceptos en textos clínicos en español con la herramienta MetaMap [155].

Reconocimiento de expresiones temporales. Una tarea muy importante dentro de la extracción de entidades nombradas en el dominio de la Medicina es la detección precisa de las expresiones temporales (fecha de aparición de una enfermedad, duración de un tratamiento, periodos de reingresos de un paciente, etc) dentro de grandes colecciones de informes clínicos. En *Lin et al.* presentaron el sistema híbrido MedTime, que combina un sistema basado en reglas lingüísticas con otro basado en AA [156]. Uno de los objetivos de este sistema híbrido fue la extracción, desde textos clínicos, de entidades relacionadas con expresiones temporales como fecha, hora, duración y frecuencia. Entre otros, los autores consideraron MetaMap para extraer características específicas del ámbito médico y los tipos semánticos, y Mallet [60] para la anotación de las expresiones temporales. Este sistema demostró que la combinación de ambas estrategias consiguió un alto rendimiento en el reconocimiento de expresiones temporales (FMeasure 87.9%).

En la investigación propuesta por *Robert et al.* presentaron un sistema híbrido para llevar a cabo el reconocimiento de eventos, expresiones temporales y relaciones temporales en textos clínicos [59]. Para llevar a cabo esta tarea los autores combinaron métodos AA con métodos basados en reglas, obteniendo buenos resultados en el reconocimiento de expresiones temporales (FMeasure 61.54%).

El poder trazar un orden cronológico automático de eventos clínicos relevantes de un paciente, podría ayudar en gran medida al profesional sanitario en la toma de importantes decisiones clínicas, como por ejemplo, en la toma de decisión del inicio/fin de un tratamiento. La correlación entre entidades puede perder sentido sin el

conocimiento de que determinados eventos clínicos ocurren en un determinado orden temporal. En el trabajo realizado por *Chang et al.* se propone un sistema híbrido llamado TEMPTING que identifica enlaces temporales entre pares de entidades [167]. El método se basa en un conjunto de informes de alta y el descubrimiento de las ocurrencias de tiempo de todos los eventos clínicos ocurridos en el transcurso del tratamiento de un paciente, entre la fecha de admisión y la fecha de alta. El método propuesto combina dos enfoques, uno basado en reglas NLP y otro basado en AA. Esta investigación ayuda al profesional sanitario en la toma de decisiones clínicas, gracias a que facilita el descubrimiento de posibles correlaciones entre entidades, como el tratamiento y los efectos adversos a medicamentos, que podrían generarse durante el tiempo de estancia del paciente en un centro sanitario.

Anonimización de datos personales. En los últimos años uno de los temas más polémicos que se plantean con la informatización de la historia clínica en los entornos sanitarios es la protección de datos de carácter personal. El preservar datos de carácter personal es una tarea imprescindible en el área de la investigación médica, donde se necesitan manejar grandes volúmenes de datos clínicos pero sin identificaciones personales que puedan afectar a la protección del paciente. La protección de los datos de carácter personal se ha convertido en un desafío para las instituciones sanitarias y los sistemas de reconocimiento de entidades nombradas pueden ayudar a paliarlo. El problema se reduciría considerable si se aplicaran técnicas de anonimización basadas en la tarea del reconocimiento de entidades nombradas, donde la información sensible pudiera ser eliminada o enmascarada (ver Figura 4.3). Este es el tema central del trabajo realizado *por Ferrández et al.* que presentan un sistema híbrido diseñado para mejorar las estrategias actuales para la "des-identificación" de nombres personales en textos clínicos [157]. Lo novedoso de este sistema se basa en la utilización combinada de los puntos más provechosos de dos

perspectivas, la basada en AA y la basada en reglas. Para la evaluación de este nuevo sistema los autores compararon su sistema NER con cinco sistemas en el campo de la extracción de entidades y la de-identificación. Los resultados demostraron una mejora de más del 26% para la métrica de F2-measure cuando la propuesta fue comparada con el mejor de los cinco sistemas analizados.

Benton et al. presentaron un sistema para eliminar números de teléfono, nombres, direcciones de correo electrónico y otros datos de identificación de los textos clínicos [158]. Para llevar a cabo esta tarea, los autores utilizaron un enfoque híbrido, basado en reglas y AA. Se usó un modelo CRF [144] para etiquetar los identificadores, mientras que dos corpus (uno basado en cáncer de mama y otro en artritis) se consideraron para entrenar y validar el modelo. Los autores valoraron su sistema frente a un conocido sistema de de-identificación, logrando un buen desempeño en la métrica recall (98.1% vs. 73.0%).

Extracción de Relaciones entre Entidades Médicas. Uno de los principales objetivos de la MT es la recuperación de conocimiento oculto en el texto. Uno de estos descubrimiento es la identificación de entidades médicas y sus relaciones, una tarea imprescindible en el ámbito de la Medicina. La extracción de relaciones consiste en encontrar las conexiones existentes entre varias entidades en un texto [160]. En el campo de la Biomedicina la extracción de relaciones ha sido ampliamente estudiada focalizándose en tareas como asociaciones gen-enfermedad, gen-fenotipo, proteína-proteína, proteína-fármaco, etc [161-163]. Gracias a las tareas de la extracción de entidades y la representación de las posibles relaciones entre las mismas se consigue que los investigadores puedan acceder a información relevante que puede ser vista desde diferentes enfoques según sus interrelaciones de forma que se puede conseguir llegar a descubrir patrones que inicialmente estaban ocultos en el texto del que se partía. *Galustian et al.* presentaron un interesante artículo donde se pone de manifiesto cómo la MT ha ayudado en la investigación contra el cáncer, con la obtención de las

relaciones entre un tipo de cáncer o fármaco y los síntomas o eficacia observada, o por ejemplo la asociación entre los distintos tipos de cáncer y un gen particular (como el AKT) [164]. Cuando se combinan las técnicas de MT con la minería de datos de microarrays [165] se pueden realizar análisis tan potentes que pueden aplicarse a la consecución de objetivos tales como cuáles son los tratamientos más efectivos para los distintos tipos de cáncer. En la investigación realizada por *Uzuner et al.* [166] se analiza la extracción y las relaciones entre distintas entidades clínicas obtenidas desde informes de alta médica. En base a los tipos semánticos definidos en el metathesaurus UMLS [195] se analizan las siguientes relaciones entre entidades: enfermedad-tratamiento, enfermedad-test, enfermedad-síntomas. Los autores emplearon un enfoque basado en ML. Por cada par de conceptos que aparecen en una oración, un clasificador SR (Semantic Relation), basado en SVM, determinará las relaciones entre ellos. El sistema se evaluó contra dos sistemas baseline, obteniendo un valor de FMeasure superior en un 15% con respecto al mejor sistema baseline.

Zhu et al. abordaron un sistema híbrido, basado en AA, diccionarios y reglas, para llevar a cabo la tarea de extraer relaciones semánticas entre conceptos médicos obtenidos desde informes de alta e informes de evolución de pacientes [168]. Se analizan tres tipos de relaciones: tratamiento-problema, pruebas-problema y problema-problema. Para la tarea automática de extracción de conceptos se utilizó un modelo de Markov. La detección de relaciones fue tratada como una tarea de categorización multiclase. Se emplearon las herramientas MetaMap [155] y Ctakes [151]. Los resultados mostraron la importancia de realizar un buen proceso de selección de características a través de diferentes fuentes externas de conocimiento, logrando así una mejora en la métrica FMeasure (74.2%) para la extracción de relaciones entre entidades médicas.

Rink et al. propusieron un sistema basado en aprendizaje supervisado para descubrir las relaciones entre problemas médicos, tratamientos y pruebas, recogidos en la historia clínica electrónica [169]. Utilizaron en su investigación varias fuentes de conocimiento externo como Wikipedia y WordNet, demostrando que la utilización de estas

herramientas externas mejoraban la tarea de extracción de relaciones. Para la extracción de las características se emplearon otros recursos terminológicos como MetaMap, UMLS [195] y Genia [196], un corpus anotado para la extracción de entidades biomédicas. La metodología empleada, como en la mayoría de trabajos relacionados, combinan CRF [144] con SVM [68] para la clasificación multiclase.

Wright et al. presentaron un sistema capaz de identificar asociaciones entre medicación, problemas y resultados de laboratorio desde un conjunto de 100,000 historias clínicas informatizadas [170]. Una de las principales ventajas que aporta esta metodología es el importante ahorro en tiempo, según los autores el procesar todo el conjunto de datos tomó alrededor de nueve minutos en la generación de miles de asociaciones entre las entidades medicación-problema y problema-laboratorio.

Otras investigaciones relacionadas con la extracción de relaciones entre entidades médicas desde textos clínicos pueden encontrarse en [171-173].

2.2.3. Clasificación diagnóstica automática.

Una de las tareas más complejas a la que diariamente se enfrenta el clínico es el abordaje del *proceso diagnóstico* donde, con ingentes cantidades de datos e información textual, provenientes principalmente de la historia clínica, el profesional sanitario debe ser capaz de inferir el mejor diagnóstico o diagnósticos posibles. Los métodos tradicionales de "búsqueda del mejor diagnóstico" son hoy en día enormemente costosos para los profesionales sanitarios debido fundamentalmente a la gran variabilidad de fuentes informacionales de las que el profesional debe nutrirse para conocer el estado de salud del paciente, el escaso tiempo que el profesional puede dedicar al paciente, la falta de recursos humanos y materiales, y otras muchas barreras que el clínico debe superar a diario. El profesional sanitario debe buscar la información más relevante, interpretar pruebas, analíticas, informes de consultas, informes de alta,

obtener conclusiones e hipótesis en base a estos resultados y establecer un diagnóstico de calidad y fiable. Frente a estos complejos factores, se hace muy difícil mantener un equilibrio perfecto entre ofrecer servicios de alta calidad, a bajo coste y con un bajo nivel de errores. Es evidente que el profesional sanitario se enfrenta diariamente con infinidad de procesos que requieren múltiples tomas de decisiones muy complejas y estresantes. *En un estudio realizado por Tudela et al., analizaron, en un total de 669 ingresos, los errores diagnósticos que se generaron en el servicio de Urgencias de un hospital de Barcelona [266]. Se detectó un 6,2% de errores diagnósticos, los errores más frecuentes se encontraron en la valoración clínica en un 42,8% y en la interpretación de los resultados de las radiografías de tórax en un 40,4%. Estos errores implicaron una consecuencia bastante importante, un retraso en el inicio del tratamiento o medida terapéutica en el 42,8% de los casos.*

Una vez abordado el complejo proceso diagnóstico, la mayoría de las instituciones sanitarias realizan, normalmente a través de las unidades de documentación clínica, una normalización "manual" de la información registrada en los distintos informes (informes de alta hospitalaria, informes quirúrgicos, etc) de la historia clínica electrónica del paciente. La codificación clínica es una tarea indispensable y de gran interés en todas las instituciones sanitarias debido a su importancia para la mejora de la gestión y calidad asistencial. Aunque el beneficio que aporta esta tarea a las instituciones sanitarias es evidente, también lleva implícita una serie de inconvenientes difíciles de abordar, es una tarea ardua con un coste elevado de tiempo, requiere un alto número de personal cualificado y al ser una tarea manual puede llevar asociada errores humanos debido a peculiaridades de la terminología médica (ambigua, no estructurada, diversa y compleja).

Ante este difícil marco, para automatizar y apoyar la difícil tarea de la codificación diagnóstica contamos con las técnicas que nos proporcionan las disciplinas de la MT, el PLN y el AA. En esta investigación, abordamos el problema de la inferencia de códigos normalizados de diagnóstico a informes clínicos textuales aplicando como base la tarea

de la clasificación de textos, analizada en la sección 2.1.4. Son muchos los estudios dedicados al desarrollo de sistemas automatizados de clasificación textual provenientes de grandes repositorios de literatura médica, como Medline o PubMed, para apoyar al profesional sanitario en la tarea de búsqueda de información de interés [258, 259], o el desarrollo de sistemas de vigilancia epidemiológica basados en informes clínicos [260], o la extracción de resultados críticos en informes radiológicos [262], etc. Sin embargo, una de las aplicaciones de la CAD, escasamente analizada hasta la fecha debido a su complejidad, así como a las peculiaridades de la terminología médica, es la asignación automática de códigos normalizados de diagnóstico a textos clínicos, denominada ***clasificación diagnóstica automática*** (CDA).

Como hemos comentado anteriormente, la CDA es un tipo de clasificación que permite inferir códigos normalizados de diagnósticos en base a información textual no estructurada, que redundará en una mayor fiabilidad, comparabilidad y usabilidad de la información clínica. Esta tarea puede proporcionar grandes ventajas ya que pueden apoyar a los profesionales sanitarios en el incremento de la calidad de la codificación, ahorrando tiempo y costes [268, 269].

La mayoría de las investigaciones existentes relacionadas con la CDA dividen los enfoques actuales con los que se aborda dicha tarea en tres grupos: enfoques basados en reglas, basados en AA y una hibridación de los dos enfoques anteriores [270]. Gran parte de estas investigaciones se han centrado en un enfoque basado en AA siguiendo un tipo de clasificación binaria o multiclase, analizaremos a continuación algunas de estas investigaciones.

En la investigación propuesta por *Pereira et al.* presentaron un sistema basado en técnicas de MT y AA para clasificar cada registro médico electrónico utilizando la codificación CIE-9 (Clasificación Internacional de Enfermedades, Novena Revisión) [271]. La investigación se basó en la clasificación CIE-9 de los diagnósticos de epilepsia encontrados en los registros electrónicos de pacientes. Se trata de un estudio sobre *clasificación multiclase*, donde los registros fueron clasificados en tres categorías de

códigos CIE. Para poner en práctica este sistema los autores emplearon varias herramientas muy utilizadas en este ámbito como son GATE, Freeling, Weka, entre otras, junto con el tesauro UMLS. Utilizaron un clasificador multiclase con el algoritmo KNN. Se obtuvieron unos resultados muy prometedores, consiguiendo un valor en la métrica FMeasure del 71,05%.

En *Pakhomov et al.* se recoge una de las pocas investigaciones que ha llevado a la práctica real un sistema de codificación automática basada en texto proveniente de las historias informatizadas [272]. Gracias a esta investigación, implantada en la Clínica Mayo, al menos el 48% de los informes clínicos electrónicos (lista de problemas iniciales), pudieron ser clasificados automáticamente con una precisión del 98.3% y un valor de FMeasure del 98.2%. El sistema denominado Autocoder en un sistema híbrido que sigue un enfoque basado en reglas y AA. El sistema se nutre de una base de registros codificados manualmente durante 10 años. Dentro del componente de clasificación basada en AA se empleó el algoritmo Naïve Bayes.

En la investigación realizada por *Schuemie et al.* propusieron un sistema de clasificación automática de información procedentes de registros electrónicos de pacientes para apoyar estudios epidemiológicos [246]. Se empleó la representación BoW y se eliminaron las negaciones encontradas en los textos clínicos (como por ejemplo, "*no focal pneumonia*"). Entre los algoritmos clasificadores empleados en este trabajo se encontraban los siguientes: Naïve Bayes, KNN, C4.5, Random Forest, RIPPER (Repeated Incremental Pruning Produce Error Repeated Incremental Pruning Produce Error R) [274], etc. La evaluación del sistema demostró que el algoritmo RIPPER obtuvo un mejor desempeño en comparación con otros algoritmos, como Naïve Bayes.

Wang et al. propusieron la creación de un sistema centrado en la Medicina China Tradicional, capaz de convertir el texto libre proveniente de registros clínicos de pacientes (no estructurados y sin normalización terminológica) en información eficaz para dar soporte y ayuda al proceso diagnóstico automático [275]. Para ello emplearon técnicas de PLN, MT y AA utilizando los clasificadores Naïve Bayes y SVM. La tarea de la

automatización diagnóstica se enfocó en esta investigación desde una perspectiva multi-clase. El sistema estaba compuesto por los siguientes módulos: módulo de análisis de los registros clínicos para la identificación de los síntomas, módulo de normalización de entidades, módulo de selección de características y módulo de entrenamiento del modelo. Los resultados de la evaluación del sistema no obtuvieron valores muy altos en Precisión y FMeasure, debido fundamentalmente, según los autores del estudio, a los complejos procesos de normalización y reconocimiento de síntomas (entidad nombrada) en los textos clínicos.

En la investigación realizada por *Sharma et al.* se presenta un sistema basado en MT con la finalidad de clasificar enfermedades dermatológicas en base a 33 atributos [276]. El trabajo se centra en la clasificación multiclase siguiendo la estrategia uno-contra-todos y la regresión logística. El sistema se focalizó en la clasificación de seis enfermedades de la piel (e.g. psoriasis). El modelo fue capaz de clasificar las enfermedades con una precisión del 94.82% y una exactitud del 98.36%.

Suzuki et al. analizaron como clasificar diagnósticos bajo el sistema DPC (Diagnosis Procedure Combination), un sistema basado en el conocido DRG (diagnosis-related group) utilizado en USA y actualmente en toda España [277]. Los códigos DPC identifican 516 enfermedades. En este estudio los autores proponen utilizar el modelo espacio vectorial para detectar códigos DPC basándose en el contenido textual de los informes de alta, para ello emplearon técnicas de MT. Se emplearon más de 15,000 informes de alta provenientes de un Hospital Universitario para identificar 14 enfermedades, entre ellas, cáncer de pulmón, diabetes, cataratas, asma, etc. En esta investigación se utilizaron dos métodos de ponderación, tf-idf y el método de entropía. Extrayendo la información relevante de los informes de alta se consiguió una correcta identificación de los diagnósticos en más del 96% de los casos.

Friedman et al. propusieron un método centrado en el mapeo de información clínica a código utilizando para ello el metatesauro UMLS [230]. Utilizaron técnicas de PLN para analizar la totalidad del documento y el sistema MedLee (Medical Language Extraction

and Encoding System), un sistema que permite extraer y codificar la información clínica textual para que pueda ser utilizado en procesos automáticos posteriores. Los resultados de la evaluación de este sistema reflejaron unos resultados muy similares a los de los expertos codificadores. Se obtuvo un valor en Precisión del 89% y un 83% en Recall.

Lussier et al. presentaron un sistema, también basado en MedLEE, para codificar la narrativa médica siguiendo la nomenclatura de SNOMED, una terminología clínica muy extendida en el mundo sanitario [226]. Otros autores desarrollaron un método para identificar pacientes con nódulos pulmonares, combinando cinco códigos de diagnóstico, cuatro códigos de procedimiento y un algoritmo PLN cuyo objetivo era buscar texto libre en informe radiológicos [206]. La sensibilidad y especificidad del algoritmo PLN propuesto ante la identificación de pacientes con presencia de nódulos fue del 96% y 86% respectivamente.

Sin embargo, aunque existen algunas investigaciones que han utilizado las técnicas de clasificación de textos convencional (binaria o multiclase), la tarea de inferencia automática del diagnóstico basada en información textual de la historia clínica no debería abordarse aplicando técnicas de clasificación tradicionales sino que debería ser resuelta desde la perspectiva de un *problema de clasificación multietiqueta*. Este paradigma ha sido el elegido en nuestra investigación para abordar la tarea CDA. La evolución del estado de salud del paciente a lo largo del tiempo hace que las patologías raramente sean únicas en una historia clínica, y por tanto, *la categorización diagnóstica no pueda ser vista como un problema de clasificación monoetiqueta sino multietiqueta*, donde un mismo elemento (informe de alta) puede ser asignado a más de una categoría (códigos diagnósticos). Un ejemplo de clasificación multietiqueta lo podemos ver representado en la Figura 2.17 donde cada documento textual, en nuestro caso cada informe de alta, puede clasificarse o codificarse utilizando múltiples diagnósticos (etiquetas), en otras palabras, cada instancia puede pertenecer a más de una clase.

Documentos	Etiquetas / Clases
Informe Alta 1	{anemia,cefalea,colesterol,diabetes}
Informe Alta 2	{colesterol,demencia,ictus}
Informe Alta 3	{nefritis,diabetes}
Informe Alta 4	{anemia,cefalea,melanoma,parkinson}

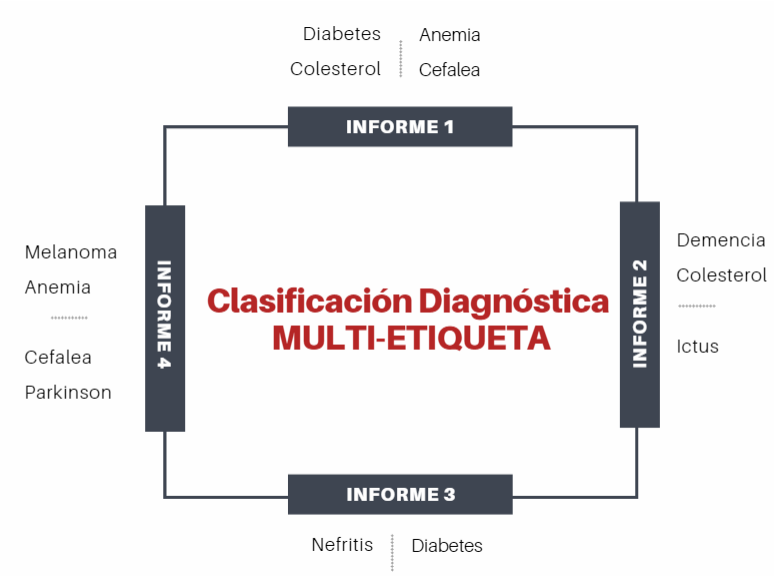


Figura 2.17. Ejemplo de Clasificación Diagnóstica Multietiqueta

*La **clasificación diagnóstica multietiqueta*** en base al texto clínico es una tarea compleja y su desarrollo y puesta en marcha, con la utilización de técnicas de AA y MT, es hoy en día muy limitada. A continuación, intentaremos profundizar en las investigaciones más destacadas sobre la clasificación diagnóstica multietiqueta, para intentar reforzar la idea del beneficio que supondría la aplicación de estas técnicas para los profesionales sanitarios.

Stanfill et al. analiza, en una revisión interesante, la dificultad real a la que se enfrentan los desarrolladores de sistemas de codificación y clasificación automática basada en información clínica [223]. De los 113 estudios incluidos en esta revisión obtuvieron una conclusión agridulce, por un lado los autores confirmaron los prometedores resultados obtenidos de estos sistemas, aunque pusieron de manifiesto el problema de la

comparabilidad debido a falta de estándares y a su aplicación en contextos muy limitados que no pueden ser reproducidos en otros entornos. Los autores observaron que cada sistema era aplicable al entorno para el que se creó, siendo difícil su expansión y aplicabilidad a otros ámbitos diferentes. Aproximadamente el 51% de los estudios compararon los resultados automáticos con los resultados obtenidos manualmente por expertos codificadores.

En los primeros desarrollos de sistemas automatizados de codificación diagnóstica multi-etiqueta existía una amplia mayoría de estudios que centraban el desarrollo de estos sistemas en la disciplina del PLN [1]. El uso de técnicas basadas exclusivamente en PLN para desarrollar sistemas de codificación automática fue perdiendo interés sobre todo en entornos tan complejos como la Medicina. Se desarrollaron nuevas técnicas para solventar las limitaciones del PLN, y comenzaron a extenderse los sistemas basados en MT [190] y AA [27]. En el trabajo desarrollado por *Metais et al.* presentaron el proyecto CIREA, un sistema cuya finalidad es automatizar la codificación ICD-10 utilizando técnicas combinadas de MT y AA [224]. El sistema intenta descubrir el diagnóstico en base al informe clínico escrito en lenguaje natural por el médico. La propuesta se centra en una clasificación multietiqueta, donde un informe médico puede ser codificado con múltiples códigos ICD. Existía una media de entre 1 y 32 diagnósticos por paciente. Se utilizaron 30,000 informes médicos con las siguientes secciones: razón del ingreso, examen de salud, evolución y conclusión. Se realizaron varias labores de normalización de todos los informes clínicos, como la aplicación del algoritmo de Porter, eliminación de tildes, armonización de caracteres especiales, etc. Los autores desarrollaron un algoritmo denominado CLO3, cuya evaluación frente al algoritmo Naïve Bayes incrementaba el resultado de la clasificación en un 6.7%. El valor de FMeasure obtenido con el algoritmo propuesto es del 76.7% frente al 70% obtenido del algoritmo Naïve Bayes. Se demostró también que la aplicación del algoritmo de Porter mejoraba el resultado de la clasificación.

En la investigación realizada por *Goldstein et al.* se evalúan tres sistemas para predecir automáticamente la codificación CIE-9-MC asociada a la información textual de informes radiológicos [268]. El primer sistema analizado está basado en Lucene[7] [249], una librería de código abierto creada en Java que permite realizar múltiples funciones de procesamiento de textos como la tokenización, eliminación de *stop-word*, etc. El segundo sistema utiliza BoosTexter [257], una aplicación AA basada en la técnica *boosting* para la categorización de textos. El tercer sistema emplea un conjunto de reglas que capta los elementos léxicos derivados de BoosTexter incluyendo información semántica del texto obtenido de cada informe radiológico. Los resultados obtenidos demostraron que el primer sistema, el basado en Lucene, arrojaba un rendimiento menor que los otros dos evaluados, obteniendo un valor en FMeasure del 66.9%. El uso del método de aprendizaje automático boosting mejoró el rendimiento del sistema pasando a arrojar un resultado en FMeasure del 80%. Aunque el sistema que arrojo un mejor desempeño fue el que combinó las técnicas de ML con un conjunto de reglas que tuvieron en cuenta las características semánticas de los informes. Los autores demostraron que la utilización de información semántica como la detección de negaciones, sinonimias y otras características específica del texto puede aumentar el rendimiento y la tarea de predecir y asignar automáticamente códigos CIE.

La pérdida de información clínica en los registros médicos electrónicos es de vital importancia debido al impacto que puede acarrear en la evaluación del proceso asistencial de un paciente. En el trabajo realizado por *Erraguntla et al.* se plantea la creación de un modelo de predicción de códigos CIE9, cuando estos no están presentes en los registros médicos electrónicos y tienen que ser basados en otros atributos textuales de los registros [90]. Las técnicas de MT fueron empleadas sobre estos datos textuales no estructurados para extraer los campos claves y posteriormente se empleó el algoritmo basado en el vecino más cercano para predecir los códigos CIE perdidos.

[7] https://lucene.apache.org/core/

Zhang et al. propusieron el desarrollo de un asistente de diagnóstico basado en el análisis de registros médicos electrónicos obstétricos bajo el enfoque de la clasificación multietiqueta [89]. La colección de informes consistió en 10,000 registros seleccionados de casos reales de 15 hospitales diferentes. Los autores consideraron tres fases diferentes: una fase de preprocesamiento (limpieza, segmentación de palabras y estandarización de datos); una fase de ingeniería de características (*latent dirichlet allocation* y modelo skip-gram) y, por último, la ejecución de una serie de métodos de clasificación multietiqueta como BP-MLL (*backpropagation multilabel learning*), RAkEL, MLkNN y CC.

Algunos investigadores, relacionados con el ámbito de la CDA, pudieron demostrar que el uso de herramientas y recursos externos que aportaban enriquecimiento semántico, como las ontologías o los tesauros médicos, podían hacer mejorar el rendimiento predictivo de la tarea de clasificación [87, 268]. De igual forma, el uso conjunto de estos recursos externos junto con herramientas que facilitaban las tareas PLN, como MetaMap, dieron muy buenos resultados en el desempeño de la clasificación diagnóstica multietiqueta. A continuación, recogeremos algunas de las escasas investigaciones relacionadas al respecto y que siguen la tendencia de la propuesta elegida en esta tesis doctoral para llevar a cabo la tarea de codificación diagnóstica multietiqueta.

Waraporn et al. propusieron la creación de un sistema para realizar la tarea de asignar códigos de diagnóstico estandarizados a registros médicos electrónicos [88]. Para tratar de mejorar esta compleja tarea, los autores propusieron un algoritmo AA llamado chiDT (*cascade hierarchical Decision Tree*) junto con la integración del enriquecimiento semántico proporcionado por una ontología sobre la cardiopatía isquémica. El uso de ontologías en sistemas que involucran tareas PLN [1] permite la estandarización y desambiguación de términos, aunque los autores no pudieron obtener resultados claramente concluyentes debido al número limitado del dataset.

Kavuluru et al. evaluaron distintos enfoques basados en aprendizaje supervisado para

llevar a cabo la tarea de asignación automática de códigos CIE-9 a los registros médicos informatizados [96]. Se empleó un dataset de 71,463 registros médicos informatizados correspondientes a informes de alta de pacientes que pasaron por un centro médico durante el periodo de dos años. Se utilizaron herramientas como MetaMap [155] y SemRep [194] para la obtención de las entidades nombradas. Se contrastaron varios métodos de transformación de problemas multilabel como BR, ECC y SVM. Para textos cortos relacionados con un subdominio particular (por ejemplo, radiología) el mejor resultado para la asignación de códigos diagnósticos se obtuvo aplicando el método CC, sin embargo, el método BR combinado con la función learning-to-rank es ideal para dominios más generales y grandes conjuntos de datos.

Yepes et al. propusieron un trabajo donde intentaron mejorar la categorización automática de citas MEDLINE con las jerarquías de códigos MeSH [202]. El problema se aborda desde la perspectiva multietiqueta, ya que un mismo descriptor puede ser asignado a más de un documento. Se emplearon dos métodos AA como son, SVM y AdaBoostM1. Los resultados demostraron la dependencia de la selección de características y los métodos AA escogidos con el rendimiento generado en la tarea de clasificación. Se compararon los resultados obtenidos con otros sistemas baseline como MTI (*Medical Text Indexer*) [197], un indexador de textos médicos. Se concluyó afirmando que la combinación de todas las características de los textos mejoraba el rendimiento de cualquier conjunto de características individuales, además la combinación del algoritmo SVM con la selección de bigramas obtuvo unos resultados superiores a los de MTI, lo que demostró que es posible mejorar este sistema de referencia.

Suominen et al. presentaron un sistema basado en clasificación multietiqueta para la asignación automática de códigos CIE9 a informes radiológicos [273]. En esta investigación demostraron la importancia de una buena selección de características para conseguir un buen desempeño del clasificador. Se combinaron técnicas del PLN con AA, partiendo de una colección de 1954 documentos con un total de 45 códigos CIE distintos

y 94 combinaciones. Este sistema empleó la herramienta MetaMap [155] y el methatesaurus UMLS [195] para llevar a cabo la fase de ingeniería de características. Se utilizaron un enfoque basado en AA con dos clasificadores en cascada (*Regularized Least Squares* y RIPPER). El resultado final del sistema propuesto arrojó un desempeño en FMeasure del 87.7%. El uso del metatesauro UMLS, el empleo de las negaciones y los sinónimos contribuyeron al aumento del rendimiento en la tarea de la clasificación. Sin embargo, uno de los inconvenientes de esta herramienta es que no parece ser extensible a otros dominios clínicos debido a las características específicas de los informes de radiología (estilo conciso y altamente dependiente del dominio).

2.2.4. Recursos y herramientas de MT orientadas al análisis textual en Medicina.

La mayoría de los sistemas basados en MT desarrollados dentro del ámbito de la Medicina deben complementarse con diferentes recursos terminológicos y herramientas, como corpus, metathesaurus, ontologías, herramientas de *tagging* y *parsing*, extractores de entidades médicas y sus relaciones, etc. A continuación, analizaremos los recursos más destacados, algunos de los cuales han sido empleados en el desarrollo del sistema propuesto en esta tesis.

Los recursos terminológicos, como los corpus, tesauros y ontologías, son esenciales en los sistemas de MT. *Los corpus* son conjuntos estructurados de textos que se clasifican y anotan según ciertos criterios lingüísticos con el fin de ser utilizados como muestra o representación de una lengua [214]. *Un tesauro* es un vocabulario controlado y estructurado que contiene términos, relaciones, jerarquías, sinónimos y otros niveles de dependencias que representan el conocimiento en un dominio determinado [39]. *Las ontologías* son herramientas que permiten que el conocimiento de un dominio específico sea organizado y representado a través de una taxonomía de conceptos,

relaciones y axiomas [66]. Algunos de los recursos terminológicos más utilizados en el dominio de la Medicina son (Tabla 2.2): **ONCOTERM [78],** un proyecto español que comenzó con la idea de proporcionar un repositorio completo de información sobre la compleja terminología asociada con el cáncer: enfermedades, medicamentos, tratamientos y otros temas relacionados. **NCBI disease corpus [210],** un corpus anotado que contiene 6,892 menciones de enfermedades que están vinculadas a 790 conceptos únicos (identificador de MeSH o identificado de OMIM [67]. **GALEN [54],** una ontología abierta dentro del dominio médico que incluye, conceptos de enfermedades, síntomas, medicamentos, procedimientos, además de las interrelaciones. **UMLS (Unified Medical Language System)** [195], uno de los metatesauro más empleados integrado por múltiples vocabularios controlados relacionados con las ciencias biomédicas y la salud.

RECURSOS TERMINOLÓGICOS	
CORPUS	
GENIA[8] [196]	**AZDC**[9] [74]
ONCOTERM[10] [78]	**NCBI**[11] **disease corpus** [210]
BIOINFER[12] [77]	
ONTOLOGÍAS	
Disease Ontology (DO)[13] [72]	**Ontology of Adverse Events** [14] [52]
GALEN[15] [54]	**Alzheimer's Disease Ontology** [16] [42]
TESAUROS	
UMLS [17] [195]	**SNOMED-CT**[18] [189]
MeSH [19] [139]	

Tabla 2.2. Recursos terminológicos en el ámbito de la Medicina

[8] http://www.nactem.ac.uk/genia
[9] http://diego.asu.edu
[10] http://www.ugr.es/~oncoterm/
[11] https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
[12] http://mars.cs.utu._/BioInfer/?q=download
[13] http://disease-ontology.org/
[14] http://www.oae-ontology.org/
[15] http://bioportal.bioontology.org/ontologies/GALEN
[16] http://bioportal.bioontology.org/ontologies/ADO
[17] http://www.nlm.nih.gov/research/umls
[18] http://www.snomed.org/snomed-ct
[19] http://www.ncbi.nlm.nih.gov/mesh

UMLS integra más de 150 vocabularios controlados como Gene Ontology (GO), MeSH, OMIM, NCBI, etc. Este tesauro incluye 3,8 millones de conceptos y 12,2 millones de nombres de conceptos distintos en la versión de 2018. **MeSH (Medical Subject Headings) [139],** un vocabulario controlado para la indexación y clasificación de información relacionada con la biomedicina y la salud. En la versión de 2019 este tesauro estaba formado por 29,351 descriptores (términos controlados).

La fase de preprocesamiento es crucial en el desarrollo de cualquier sistema de MT, en ella el conjunto de documentos textuales se transforma en un conjunto de información estructurada mediante la aplicación de una serie de técnicas como la eliminación de palabras vacias, tokenización, etiquetado de palabras, etc. Para automatizar estas complejas tareas existen ***herramientas que facilitan el preprocesamiento de las colecciones textuales***, algunas de ellas se recogen en la Tabla 2.3.

PREPROCESAMIENTO DE COLECCIONES TEXTUALES	
GENIA sentence splitter[20][16]	**GENIA tagger**[21][16]
Stanford Part-of-Speech Tagger[22][43]	**Stanford Parser**[23] [51]
Enju[24] [53]	**TreeTagger**[25] [65]
FreeLing[26] [75]	**SVMTools**[27] [76]

Tabla 2.3. Herramientas para el preprocesamiento de colecciones textuales

La ***detección y extracción de entidades nombradas y la identificación de sus interrelaciones*** desde colecciones clínicas textuales, juegan un papel fundamental en el dominio de la Medicina. Algunas de las más importantes herramientas para realizar

[20] http://www.nactem.ac.uk/y-matsu/geniass/
[21] http://www.nactem.ac.uk/GENIA/tagger/
[22] https://nlp.stanford.edu/software/tagger.shtml
[23] https://nlp.stanford.edu/software/lex-parser.shtml
[24] http://nactem.ac.uk/enju/
[25] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[26] http://nlp.lsi.upc.edu/freeling/
[27] http://www.lsi.upc.edu/~nlp/SVMTool/

estas funciones y que son ampliamente utilizadas en el ámbito médico se recogen en la Tabla 2.4.

EXTRACCIÓN DE ENTIDADES NOMBRADAS Y RELACIONES	
FACTA[28] [99]	**DNorm**[29] [111]
PolySearch[30] [267]	**Chilibot**[31] [82]
MEDIE[32] [265]	**BioIE**[33] [264]

Tabla 2.4. Herramientas para la extracción de entidades nombradas y sus relaciones

Algunas de las tareas típicas que engloba la mayoría de los sistemas de MT, como la tokenización, lematización, reconocimiento y extracción de entidades nombradas, etc, pueden ser llevadas a cabo por herramientas que integren todos estos recursos y unifiquen estas funcionalidades en un único sistema. Algunas de estos ***sistemas integrales,*** más utilizados en las investigaciones de MT y Medicina, son, entre otros, MetaMap, RapidMiner, Weka, Mulan, Meka, Gate, @Note y UIMA (ver Tabla 2.5). **MetaMap** [155, 350] es una de las herramientas más utilizadas en el ámbito de la Medicina, su principal finalidad es encontrar conceptos relevantes desde colecciones de textos utilizando como base terminológica y semántica el metatesauro UMLS. MetaMap emplea, entre otras, técnicas de PLN. Realiza múltiples tareas como la interpretación semántica y desambiguación de textos clínicos, detección de acrónimos, detección de negaciones, etc. En el Capítulo 3, analizaremos con más detalle las funcionalidades de MetaMap que han sido utilizadas en esta tesis doctoral. **GATE (General Architecture for Text Engineering)** [261], es un software libre de código abierto muy utilizado por todos los profesionales dedicados al procesamiento de textos.

[28] http://www.nactem.ac.uk/facta/
[29] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/DNorm.html
[30] http://polysearch.cs.ualberta.ca/index
[31] http://www.chilibot.net/
[32] http://www.nactem.ac.uk/medie/
[33] http://bioie.biopathway.org/

HERRAMIENTAS QUE INTEGRAN TAREAS DE MINERÍA DE TEXTOS	
MetaMap[34] [155,263]	**GATE (General Architecture for Text Engineering)**[35] [261]
UIMA (Unstructured Information Management Architecture)[36] [256]	**Apache cTAKES (clinical Text Analysis and Knowledge Extraction System)**[37] [151]
WEKA (Waikato Environment for Knowledge Analysis)[38] [254]	**MEKA**[39][253]
Mulan[40][252]	**KNIME (Konstanz Information Miner)**[41] [251]
@Note[42] [250]	**RapidMiner**[43] [255]

Tabla 2.5. Herramientas integrales para análisis de colecciones textuales

La adaptación de GATE al ámbito de la Medicina, con la incorporación de múltiples *plugins*, ha sido muy importante para el avance en numerosas investigaciones en salud. Algunos de los recursos que incluye GATE son, ABNER, MetaMap, GENIA, Penn BioTagger, Linked Life Data, OrganismTagger, etc. **UIMA (Unstructured Information Management Architecture)** [256] es un software, con grandes similitudes a GATE, cuyo principal objetivo es analizar grandes volúmenes de información no estructurados para descubrir conocimiento relevante para el usuario. Entre algunas de sus muchas funcionalidades están la detección de entidades y sus relaciones, anotación, sumarización de documentos, analizadores gramaticales, análisis multilingües, etc.

[34] https://metamap.nlm.nih.gov/
[35] https://gate.ac.uk/
[36] https://uima.apache.org/external-resources.html
[37] http://ctakes.apache.org/
[38] http://www.cs.waikato.ac.nz/ml/weka/
[39] http://meka.sourceforge.net/
[40] http://mulan.sourceforge.net/
[41] https://www.knime.com/
[42] http://anote-project.org/
[43] https://rapidminer.com/products/studio/

Apache cTAKES (clinical Text Analysis and Knowledge Extraction System) [151] es un sistema de código abierto específicamente diseñado para la extracción de información relevante desde registros médicos electrónicos. Dispone de una gran variedad de funcionalidades para el análisis de textos y la extracción de información en Medicina: detección de negación, MER, anotación de mención de fármacos, detección de correferencias, etc. **WEKA (Waikato Environment for Knowledge Analysis)** [254] es una herramienta de código abierto ampliamente utilizada por los investigadores en el descubrimiento de conocimiento. Permite realizar distintas tareas relacionadas con el procesamiento de textos como tokenización, eliminación de stop-word, *stemmin*g, representación vectorial, clasificación, clustering, etc. Otras herramientas que realizan similares funcionalidades bajo el enfoque del aprendizaje multi-etiqueta son **MEKA** [253] **y Mulan** [252]. **KNIME (Konstanz Information Miner)** [251] es una plataforma de código abierto que incorpora múltiples funcionalidades para realizar tareas de MT cómo POS tagging, *stemming* , representación de documentos (BOW), clasificación de documentos, etc. **@Note** [250] es una novedosa herramienta basada en MT exclusivamente diseñada para el dominio de la Biomedicina. Es capaz de realizar las principales tareas de recuperación y extracción de información, permite realizar tareas de preparación de colecciones textuales a través de su módulo de MT, incorpora un módulo para el reconocimiento de entidades nombradas (NER), visualiza las principales estadísticas de los diferentes corpus integrados e incorpora características de GATE a través de plugins. **RapidMiner** [255] es una herramienta que permite realizar múltiples tareas de la MT y AA. Permite realizar tareas de pre-procesamiento de conjuntos textuales, distintos métodos de selección y extracción de características, transformación de atributos, modelos predictivos, validación de modelos, etc.

3

MiNerDoc

Actualmente, en la práctica clínica diaria existen escasas herramientas que extraigan conocimiento y den valor a la información textual contenida en los registros electrónicos de salud. La mayoría de los sistemas de información de los hospitales y centros de salud carecen de herramientas que permitan facilitar el trabajo de los profesionales sanitarios para realizar tareas complejas encaminadas a la labor preventiva. Para paliar este problema hemos comprobado que la aplicación de técnicas de la Minería de Textos (MT), mediante el reconocimiento de entidades médicas y la clasificación multietiqueta, pueden apoyar tareas del ámbito médico, como la detección de factores de riesgo y la inferencia diagnóstica. Sin embargo, hemos comprobado, en base a la literatura analizada, que los sistemas de MT existentes para facilitar la toma de decisiones clínicas son insuficientes, suelen desarrollar una única tarea en un área específica de conocimiento, son poco amigables para usuarios no expertos en análisis textuales y son de difícil aplicación en distintos dominios para los que fueron desarrollados. Para abordar este problema, proponemos la creación de un novedoso sistema de MT, llamado MiNerDoc, cuyo objetivo principal es apoyar el proceso de toma de decisiones clínicas mediante el análisis de informes clínicos textuales. En este capítulo analizaremos en profundidad el sistema propuesto, describiremos su arquitectura, los requerimientos y recursos software empleados para su desarrollo, la metodología aplicada para llevar a cabo las dos tareas principales que lo integran (Reconocimiento de Entidades Médicas (MER, por sus siglas en inglés) y Clasificación Diagnóstica Automática (CDA)) y por último, detallaremos sus principales funcionalidades.

3.1. Descripción general y arquitectura

MiNerDoc es una aplicación de escritorio implementada en Java cuya principal finalidad es inferir conocimiento desde textos clínicos en inglés. Hemos desarrollado un sistema, basado en MT, que interpreta el contenido textual proveniente de una colección de informes de altas e infiere el conocimiento necesario para apoyar la toma de decisiones clínicas. Se basa en dos tareas principales, en primer lugar, permite identificar factores de riesgo en base a la detección automática de entidades médicas y en segundo lugar, permite inferir códigos de diagnóstico normalizados realizando una predicción automática de la categoría o categorías diagnósticas a la que pertenecen dichos informes (ver Figura 3.1).

El sistema desarrollado en esta tesis está compuesto por dos subsistemas principales (ver Figura 3.2 y 3.3) que conforman su arquitectura general. El primer subsistema realiza la detección automática de factores de riesgo en base al reconocimiento de cinco entidades médicas siguiendo un enfoque basado en diccionarios (ver Figura 3.2).

Figura 3.1. Esquema básico del sistema TM propuesto (MiNerDoc)

El segundo subsistema realiza la predicción de códigos de diagnósticos normalizados siguiendo un método híbrido que combina un enfoque basado en AA y un enfoque basado en diccionarios (ver Figura 3.2). Una de las características más importantes de MiNerDoc es la utilización de dos recursos, ampliamente utilizados en el campo de la Medicina, como son la herramienta MetaMap y el metatesauro UMLS. Ambos recursos han sido imprescindibles para el desarrollo de los dos subsistemas que componen MiNerDoc.

A continuación, analizaremos la arquitectura de MiNerDoc, analizando las entradas y salidas a los dos principales subsistemas que lo integran, el subsistema MER y el subsistema desarrollado para realizar la clasificación diagnóstica.

Sistema MER. El primer subsistema de MiNerDoc incluye un sistema MER que permite detectar cinco entidades médicas (enfermedad, farmacología, región/parte del cuerpo, procedimiento/prueba y hallazgo/síntomas). Este subsistema es responsable de la detección de diferentes factores de riesgo o alertas clínicas de interés para el clínico en base a las entidades médicas extraídas del texto clínico. Los datos de entrada y salida del primer subsistema de MiNerDoc son los siguientes:

- **Información de entrada al sistema MER de MiNerDoc.** Existen dos flujos de entrada de información al sistema MER de MiNerDoc que pueden realizarse de forma independiente: i) el sistema requiere que el usuario defina en lenguaje natural, una única vez al inicio o cuando se necesite ampliar, los principales factores de riesgo asociados a un área de riesgo específica (enfermedad cardíaca o respiratoria); ii) el sistema permitirá al usuario cargar uno o varios informes clínicos provenientes de fuentes externas (formato texto) o crear un informe directamente desde el editor de textos de la aplicación.

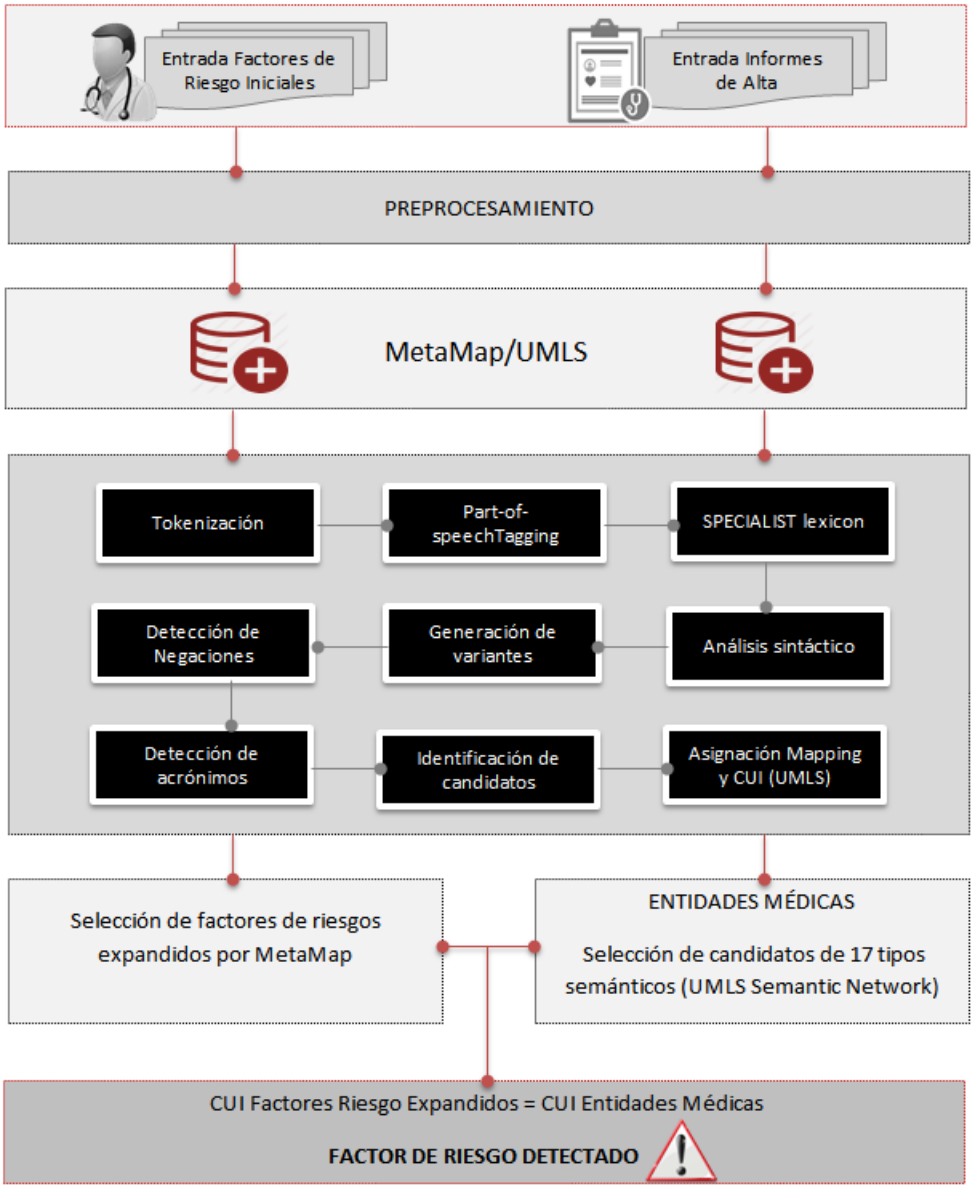


Figura 3.2. Arquitectura sistema MER de MiNerDoc para la detección de factores de riesgo.

- **Información de salida del sistema MER de MiNerDoc.** La información de salida generada por el sistema MER de MiNerDoc son dos: i) las entidades médicas reconocidas en los informes clínicos de entrada clasificados en 5 grupos y; ii) los factores de riesgo expandidos por MetaMap en base a los factores de alerta

iniciales introducidos por el usuario en lenguaje natural. De estas dos salidas se obtiene la información de mayor interés que son los factores de riesgo detectados automáticamente en base a los dos puntos mencionados.

El flujo de procesos que sigue el sistema MER de MiNerDoc (Figura 3.2) es el siguiente, los informes de alta recibidos por MiNerDoc, son normalizados automáticamente mediante la aplicación de diferentes procesos aplicando varios scripts, como la eliminación de caracteres no ASCII, borrado de líneas en blanco, etc. A continuación, a través de la herramienta MetaMap, se aplican una serie de fases PLN (por ejemplo, tokenización, etiquetado o desambiguación) para cada informe clínico, con el objetivo de obtener un conjunto de características de alta calidad. El informe clínico inicial (o informes) se fragmenta en oraciones o frases simples y se segmentan en términos candidatos a través del servidor de etiquetado MetaMap (MedPost / SKR POS). Gracias al servidor de desambiguación de MetaMap (WSD), el término candidato más apropiado se elige según el contexto desde el que se obtuvo. En base a UMLS y su red semántica, serán seleccionados diferentes tipos semánticos que permitirán clasificar cada término candidato extraído. Esta fase es responsable de identificar cinco entidades médicas en base a 17 tipos semánticos UMLS (por ejemplo, enfermedad o síndrome, signo o síntoma, hallazgo, fármaco, etc.). Finalmente, gracias a MetaMap, MiNerDoc expande los factores de alerta que inicialmente fueron proporcionados por el usuario y añadirá todas las variantes terminológicas posibles (sinónimos, derivaciones lingüísticas o interpretación de acrónimos). Los factores de riesgo expandidos se comparan con las entidades médicas extraídas en la fase anterior, en caso de que se produzca una coincidencia de acuerdo con los identificadores únicos de conceptos de UMLS (CUI) se detectará un factor de riesgo que será presentado al usuario en una ventana emergente de alerta. Estos procesos serán analizados con mayor detalle en la Sección 3.3 (Metodología).

Sistema CDA. El objetivo principal del segundo subsistema de MiNerDoc es la asignación automática de códigos de diagnóstico estandarizados (códigos MeSH) de acuerdo al contenido textual de los informes clínicos. Este segundo subsistema tiene los siguientes flujos de entrada y salida:

- **Información de entrada al sistema CDA de MiNerDoc.** La entrada al sistema que permite la inferencia diagnóstica está formada por los nuevos informes de alta sin codificar que necesitan ser categorizados, estos informes podrán ser cargados desde fuentes externas o podrán ser creados directamente desde el editor de textos de la aplicación. A estos informes se le aplicará el modelo de aprendizaje que utiliza MiNerDoc y que ha sido entrenado con un conjunto de 1,210 informes de alta, de naturaleza multietiqueta, provenientes de la colección MIMIC [55].
- **Información de salida al sistema CDA de MiNerDoc.** Los datos de salida que proporciona el sistema de clasificación de MiNerDoc son: i) predicción de grupos de diagnóstico MeSH (22 categorías) de uno o varios informes de alta; ii) representación gráfica del top ten de términos candidatos que han formado parte del proceso de clasificación (mayor puntuación de ranking MMI de MetaMap) para la inferencia diagnóstica de un único informe de alta; iii) representación gráfica de la predicción diagnóstica por categorías MeSH de un conjunto de informes de alta, y por último; iv) los ficheros arff generados automáticamente (test y training) en base a los datos de entrada, útiles para realizar otros análisis externos con los datos generados.

El flujo de procesos que sigue el sistema CDA de MiNerDoc (Figura 3.3) es el siguiente, el sistema parte de la creación de nuevos informes clínico desde el editor de textos de MiNerDoc o se parte de informes de alta previamente creados. Los informes clínicos se normalizan y limpian automáticamente. A continuación, un elemento esencial de esta

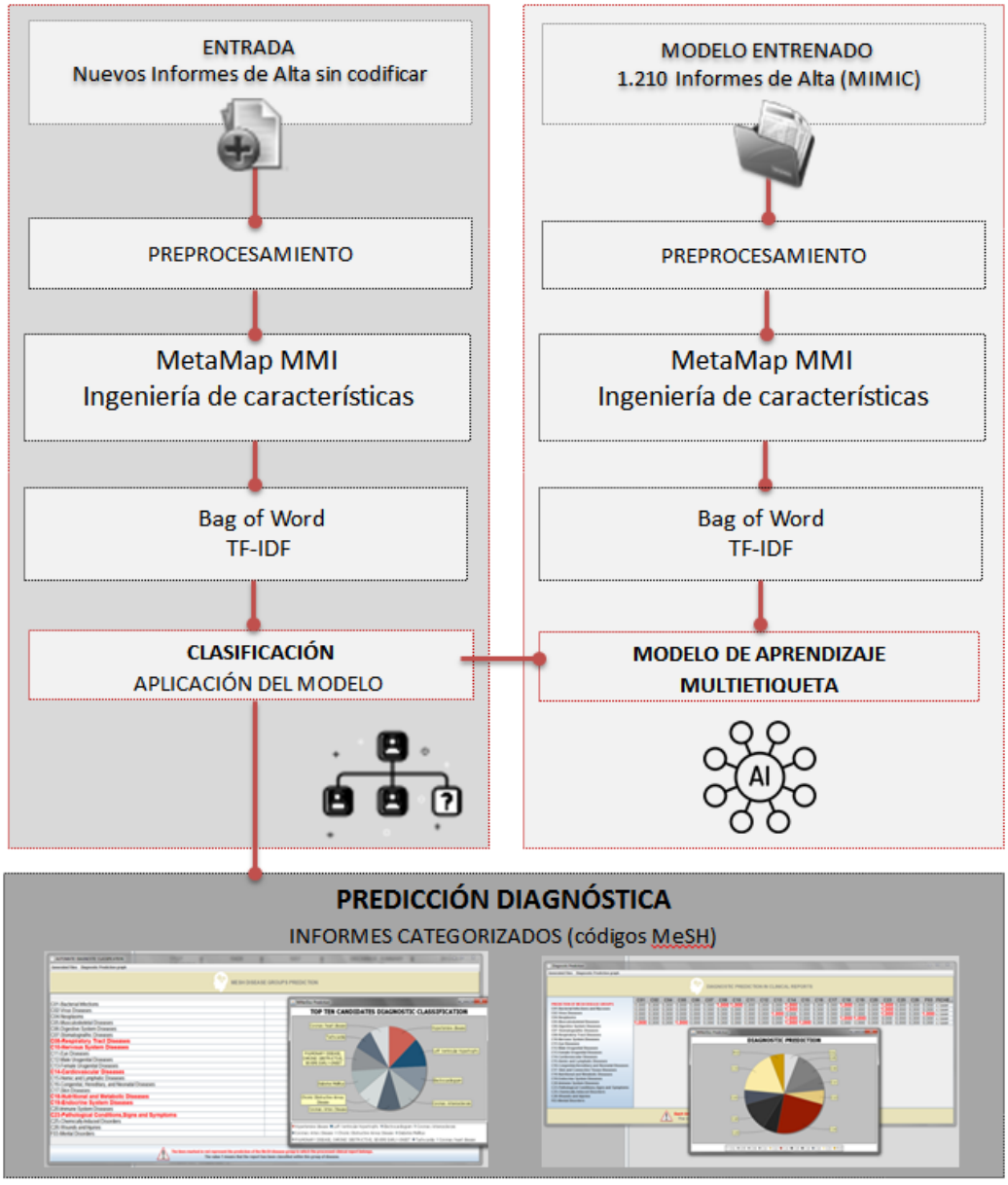


Figura 3.3. Arquitectura sistema CDA de MiNerDoc (clasificación multietiqueta).

fase es el uso del método de indexación MetaMap (MMI), este proceso permite extraer un conjunto de términos candidatos adecuados de cada informe clínico. Una vez obtenidos este conjunto de candidatos, el método MMI realiza un ranking de los

candidatos a través de una función denominada función de ranking MMI[44]. Para tener una mayor expansión de términos, hemos seleccionado todos los términos candidatos mapeados a través del método MMI. Posteriormente, se aplican diferentes técnicas de MT para obtener un conjunto de características de alta calidad en base al contenido textual de los informes clínicos. Algunas de las etapas más importantes llevadas a cabo en esta fase son: i) *selección de 5 tipos semánticos UMLS* - gracias a UMLS y la red semántica (*Disease or syndrome, Injury or poisoning, Neoplastic process, Finding y Mental or Behavioral Dysfunction*; ii) *detección automática de negaciones y*; iii) *post-procesamiento de características.* A continuación, los informes clínicos, preprocesados y estandarizados en las fases anteriores, se transforman en un modelo adecuado para un análisis posterior eficiente (modelo BoW y peso TF-IDF). Por último, en la fase de descubrimiento, MiNerDoc aplicará el modelo entrenado con los 1,210 informes de alta etiquetados para obtener la predicción diagnóstica final. En la Sección 3.3 (Metodología) serán analizados detalladamente estos procesos.

3.2. Requerimientos y recursos software empleados.

Es importante destacar que la instalación de MiNerDoc necesita unos requerimientos mínimos que son realmente simples: i) instalación de MetaMap y su API de Java; ii) aceptación del acuerdo de licencia UMLS; iii) ejecución de los servidores de MetaMap, servidor *Part-of-Speech Tagger SKR/Medpost* y servidor de desambiguación *Word Sense Disambiguation*, y por último; iv) un gestor de bases de datos MySQL.

A continuación detallaremos los principales recursos software que han sido imprescindibles para el desarrollo del sistema propuesto en esta tesis doctoral:

44 https://ii.nlm.nih.gov/MTI/Details/mmi.shtml

MetaMap. MetaMap [155] es un software desarrollado por el Dr. Alan Aronson de la National Library of Medicine y es considerada como una de las herramientas más utilizadas en el ámbito de la Medicina [146-150, 248]. Su principal finalidad es encontrar conceptos relevantes desde colecciones de textos utilizando como base terminológica y semántica el metatesauro UMLS. UMLS [195] es un sistema que integra más de 150 fuentes de vocabularios y clasificaciones (SNOMED, ICD-10-CM, MeSH, etc), con más de tres millones de conceptos centrados en el ámbito biomédico. MetaMap emplea, entre otras, técnicas de PLN, realiza múltiples tareas como la detección y desambiguación de entidades médicas, detección de acrónimos y detección de negaciones. MetaMap incorpora enriquecimiento semántico ya que permite extraer 133 tipos semánticos, como por ejemplo las categorías "*Disease or Syndrome*" o "*Diagnostic Procedure*". MetaMap es altamente configurable. Para la generación de candidatos a términos, dentro de las múltiples opciones de configuración que ofrece la herramienta, MetaMap ofrece una funcionalidad que es el método *MetaMap Indexing* (MMI) [240] que permite extraer, a partir de un documento textual, una lista global de los términos candidatos ordenados por una puntuación. Cuanto mayor es este valor, mayor es la relevancia del concepto UMLS dentro del texto. Este valor denominado "Meta Mapping" [263], es una medida aplicada a cada término candidato y mide el grado de similitud o exactitud de dicho candidato con el concepto encontrado en el metatesauro UMLS.

MiNerDoc se nutre directamente de la herramienta MetaMap y su uso está presente en los dos sistemas que lo integran, sistema MER y sistema CDA. Por un lado, MiNerDoc utiliza la red semántica del metatesauro UMLS (incluido en MetaMap) a través de la selección de determinados tipos semánticos que permiten centrar las categorías de mayor interés en el ámbito clínico (diagnóstico, fármacos, hallazgos, etc). Utilizando UMLS, MetaMap permitir llevar a cabo una conceptualización semántica de los textos clínicos, hecho que posibilita la captación de características de mayor calidad. La red semántica de UMLS ha sido utilizada en las dos funcionalidades principales de MiNerDoc

empleando 17 tipos semánticos distintos en el sistema MER (ver Sección 3.3.1) y 5 tipos semánticos para el sistema CDA (ver Sección 3.3.2). La elección de los tipos semánticos elegidos en el sistema MER se basó en la importancia de identificar los grupos de entidades médicas de mayor interés en la detección de los factores de riesgos, siempre basándonos en información textual. Así, se seleccionaron 5 grupos de entidades (*disease, pharmacologic, región/part body, Procedure/test* y *Finding/Sign*) que permiten identificar claramente un amplio espectro de factores de riesgo (enfermedades concomitantes, interacciones de fármacos, factores relacionados con el tratamiento, hallazgos clínicos, etc). En cuanto al sistema CDA de MiNerDoc, se seleccionaron 5 tipos semánticos al ser los que mejor identificaban y centraban la categoría diagnóstica. De este modo, se generan características focalizadas exclusivamente en el diagnóstico que consiguen mayor calidad y menor ruido en los *datasets* utilizados para llevar a cabo la clasificación diagnóstica.

Además de nutrirse de UMLS, MiNerDoc utiliza el método MMI [240] en su sistema CDA (ver Sección 3.3.2). Esta utilidad permite que MiNerDoc pueda procesar la información textual de cada informe clínico analizado (a través de varias técnicas PLN) permitiendo extraer una serie de términos candidatos que serán ordenados en base a un ranking de concordancia con el metatesauro UMLS. La función MMI de MetaMap, además de permitir obtener una conceptualización diagnóstica de los informes clínicos procesados, permite realizar de forma más ágil y eficaz la fase de ingeniería de características. De este modo, partiendo de la lista de términos candidatos obtenidos, gracias a MMI, será posible que MiNerDoc realice la gestión de algunos elementos esenciales en el ámbito clínico como son la detección de negaciones, la desambiguación, la detección de acrónimos, la selección de determinados tipos semánticos, etc.

MetaMap API. MetaMap Java Api permite el uso del motor de MetaMap para la utilización de todas sus funcionalidades desde aplicaciones Java[45]. El motor de mapeo de MetaMap está escrito en SICStus Prolog. Con el fin de facilitar su utilización en aplicaciones Java el sistema usa PrologBeans para permitir un mejor acoplamiento entre la API y el motor de mapeo. Es necesaria la instalación completa del paquete MetaMap y la versión de Java 1.6 SDK o posterior. Como requerimiento para el uso de la API es necesario que el servidor de MetaMap, MedPost/SKR POS Tagger y el servidor de desambiguación de palabras estén en ejecución, para ellos debemos utilizar las siguientes opciones:

MetaMap Server (mmserver)	>>	*mmserver12*
Word Sense Disambiguation (WSD) Server	>>	*wsdserverctl_start*
Part of speech tagger de MetaMap	>>	*skrmedpostctl_start*

La desambiguación es un tema importante de estudio del área del PLN. La resolución de la ambigüedad de las palabras en un contexto determinado es una tarea bastante compleja y aún más en el ámbito de la Medicina. En el sistema propuesto, hemos empleado el uso del desambiguador de MetaMap, para determinar el sentido correcto de una palabra en un contexto dado y además nos permitirá expandir los rasgos de las colecciones. El servidor de desambiguación que ofrece MetaMap es una herramienta muy útil dentro del dominio médico. En los casos que MetaMap haya asignado dos o más conceptos a una entidad del texto, el servidor WSD intentará determinar qué concepto es la mejor opción para esta entidad utilizando para ello el contexto donde se encuentra el término analizado. También es posible realizar la parametrización del uso del algoritmo Negex de MetaMap para realizar la detección de negaciones en un texto clínico.

45 https://metamap.nlm.nih.gov/JavaApi.shtml

La API de MetaMap es utilizada en MiNerDoc para establecer comunicación entre nuestro sistema (implementado en Java) con las funcionalidades de MetaMap. A través de esta API, MiNerDoc permite realizar algunas de las siguientes tareas:

- Interpretación semántica de documentos clínicos.
- Desambiguación terminológica
- Detección e interpretación de acrónimos
- Expansión de vocabulario centrado en el ámbito médico (mayor riqueza terminológica)
- Detección de negaciones (conceptos negados)
- Aplicación de tareas PLN para el preprocesamiento de informes clínicos (segmentación de textos, etiquetado gramatical, generación de variantes lingüísticas, etc)
- Selección de tipos semánticos UMLS (categorización semántica)

Mulan y Meka. Estas dos herramientas han sido importantes para el desarrollo de MiNerDoc. ***Mulan*** [210] es una librería basada en Weka que utiliza técnicas de aprendizaje automático para tratar conjuntos de datos multietiqueta. Esta librería es de código abierto y está realizada en Java. Actualmente no existen ninguna GUI para su uso aunque existe una API para que pueda ser utilizada desde cualquier programa. Esta librería ha sido utilizada para implementar la fase experimental de la tarea CDA del sistema propuesto en esta tesis doctoral (ver Capítulo 5). ***Meka*** [209] está basado en el framework Weka y proporciona apoyo para el desarrollo, ejecución y evaluación de clasificadores multietiqueta. Al igual que Weka también está escrito en Java. Meka puede ser utilizado desde su interfaz, desde línea de comando o puede ser integrado en aplicaciones más complejas a través de su API.

Hemos desarrollado varios shell scripts para que desde la aplicación MiNerDoc puedan ejecutarse funcionalidades de Meka y poder realizar así los siguientes procesos:

- Realizar la normalización de las colecciones textuales con la aplicación de técnicas de MT como la aplicación de la técnica *stemming* (algoritmo de Porter).
- Realizar la tokenización específica empleada para la construcción de los distintos *datasets* (unigramas y bigramas).
- Eliminación de *stop-words* (aplicando una lista de creación propia).
- Representación de documentos (informes de alta) basándonos en la función de ponderación tf-idf.
- Creación del modelo de clasificación que vamos a utilizar en la tarea de clasificación diagnóstica multietiqueta.
- Aplicación del modelo de clasificación para obtener resultado de predicción de la codificación diagnóstica.

MeSH Browser. Para realizar la tarea de clasificación diagnóstica, MiNerDoc sigue el enfoque de la clasificación supervisada. Este enfoque requiere partir de una colección previamente categorizada, para ello cada informe de alta de la colección de partida fue codificado manualmente por un médico experto en documentación clínica del Hospital Universitario Reina Sofía con uno o varios descriptores MeSH de enfermedad. MeSH (*Medical Subject Headings*) [139] es un sistema de vocabulario controlado que se emplea para indexar, catalogar y buscar artículos científicos. MeSH contiene una serie de descriptores organizados en 16 categorías (ver Tabla 3.1). Para codificar nuestra colección original seleccionamos *22 jerarquías diagnósticas* procedentes de dos categorías MeSH, *"C-Diseases*" y "*F-Psychiatry and Psychology*", por tratarse de los grupos de enfermedad predominantes en la colección MIMIC [55]. Las 22 clases que han servido de guía para codificar los informes de altas han sido las recogidas en la Tabla 3.2. El buscador MeSH es un sistema online de búsqueda de vocabulario diseñado para ayudar a localizar rápidamente conceptos o términos que son mostrados según una jerarquía. Esta herramienta ha servido de apoyo para realizar la clasificación previa de los informes de alta que conforman nuestra colección inicial.

JERARQUÍAS MESH	
A	Anatomy
B	Organisms
C	Diseases
D	Chemicals and Drugs
E	Analytical, Diagnostic, and Therapeutic Techniques and Equipment
F	Psychiatry and Psychology
G	Phenomena and Processes
H	Disciplines and Occupations
I	Anthropology, Education, Sociology and Social Phenomenon
J	Technology, Industry, and Agriculture
K	Humanities
L	Information Science
M	Named Groups
N	Health Care
V	Publication Characteristics
Z	Geographicals

Tabla 3.1. Jerarquías MESH

CLASES MESH	
C01	Bacterial Infections and Mycoses
C02	Virus Diseases
C04	Neoplasms
C05	Musculoskeletal Diseases
C06	Digestive System Diseases
C07	Stomatognathic Diseases
C08	Respiratory Tract Diseases
C10	Nervous System Diseases
C11	Eye Diseases
C12	Male Urogenital Diseases
C13	Female Urogenital Diseases and Pregnancy Complications
C14	Cardiovascular Diseases
C15	Hemic and Lymphatic Diseases
C16	Congenital, Hereditary, and Neonatal Diseases and Abnormal.
C17	Skin and Connective Tissue Diseases
C18	Nutritional and Metabolic Diseases
C19	Endocrine System Diseases
C20	Immune System Diseases
C23	Pathological Conditions, Signs and Symptoms
C25	Chemically-Induced Disorders
C26	Wounds and Injuries
F03	Mental Disorders

Tabla 3.2. Descriptores MESH seleccionadas para codificar la colección de informes de altas

Un ejemplo de salida del buscador MesH se observa en la Figura 3.4, donde después de la introducción de un diagnóstico concreto (“*Myocardial Infarction*”) aparecerá la jerarquía MeSH y el descriptor donde se clasifica (“Cardiovascular Diseases -C14 “).

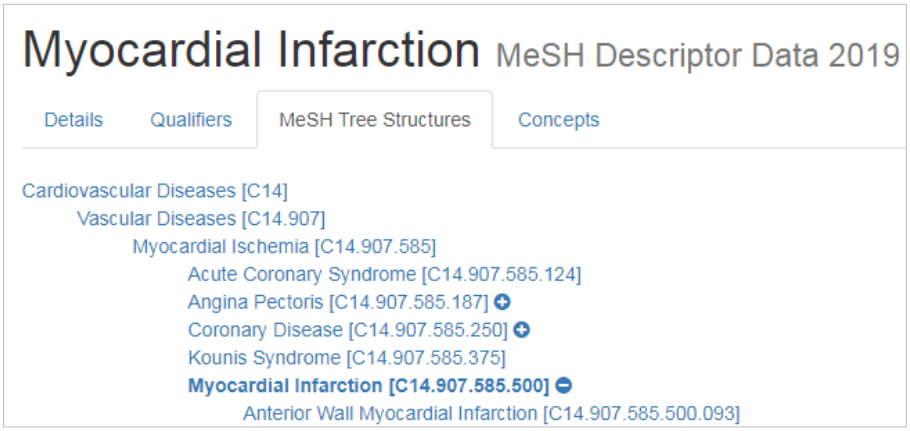


Figura 3.4. Ejemplo salida buscador MeSH

3.3. Metodología

En esta sección abordaremos en profundidad las metodologías llevadas a cabo para realizar las dos tareas principales de MiNerdoc, por un lado hemos desarrollado una metodología para la extracción de entidades médicas relevantes que nos llevará a la detección de factores de riesgo y en segundo lugar hemos creado una metodología, denominado dCSE (*diagnostic Classification with Semantic Enrichement*), para llevar a cabo la predicción automática de una o varias categorías normalizadas de diagnóstico. Ambas metodologías tiene su base en el análisis de informes clínicos textuales en inglés para conseguir construir una herramienta de apoyo en la toma de decisiones clínicas. La metodología seguida para construir el sistema MER se basa en un enfoque basada en diccionarios (MetaMap y UMLS) y para construir la metodología dCSE hemos seguido un enfoque híbrido, donde se ha unificado una solución basada en AA con una basada en diccionarios.

3.3.1. Metodología propuesta para el reconocimiento de entidades médicas y detección de factores de riesgo.

La primera tarea a desarrollar fue la creación de un sistema MER para su posterior aplicación en la detección de factores de riesgo o alertas clínicas en base a informes clínicos textuales. Como hemos podido ver en la Sección 2.1.4, existen varios enfoques para afrontar la tarea MER, nuestra propuesta se basa en un enfoque basado en diccionarios. Partiremos de la *combinación de técnicas de MT* [190] *y el uso del metatesauro UMLS* [195], basándonos para ello en una las herramientas más utilizadas en el ámbito médico como es MetaMap [155]. Nuestro sistema MER permitirá la detección y extracción desde un informe de alta de las siguientes entidades médicas: *diagnóstico, farmacología, procedimientos o test diagnósticos, hallazgos o síntomas y localización anatómica.* Una funcionalidad importante de nuestro sistema MER es la capacidad de detectar y extraer negaciones. En el dominio médico es de especial relevancia tener presente la detección de negaciones del tipo "*no evidence of pulmonary embolus*" o "*denied any chest pain*". El esquema metodológico de nuestro sistema MER puede verse en la Figura 3.5, a continuación detallaremos las fases que sigue dicho sistema:

Fase de preprocesamiento. La fase de preprocesamiento de una colección textual es una de las fases más laboriosas e importantes en la construcción de un sistema basado en MT. Para llevar a cabo una adecuada selección de rasgos se requiere realizar un proceso previo de transformación de los informes de alta originales. El punto de partida de esta fase será un informe de alta o un documento clínico de contenido textual, que puede contener información no estructurada, no normalizada y descrita en lenguaje natural. Este documento inicial es sometido a tareas de preprocesamiento para que pueda ser analizado posteriormente por la herramienta MetaMap [155]. Para ello se realizaran, entre otras, las siguientes tareas de "preparación textual":

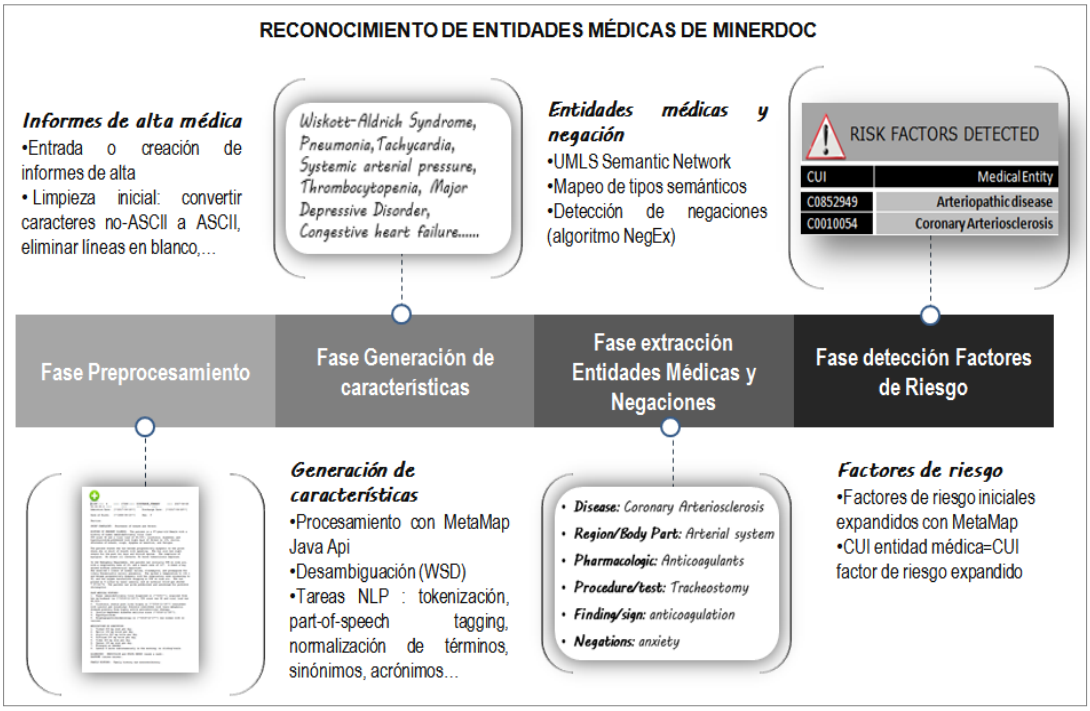


Figura 3.5 Metodología aplicada en el sistema MER de MiNerDoc

- Eliminación de caracteres especiales, ya que el software que posteriormente usaremos en las siguientes fases (MetaMap) sólo acepta texto ASCII. La existencia de caracteres no ASCII hacen que MetaMap no realice correctamente el procesamiento de los informes.
- Eliminación de líneas en blanco en los informes clínicos, necesario para que la siguiente fase puede ejecutarse sin problemas.

Para realizar la automatización de la fase de preprocesamiento de los informes de alta de entrada al sistema MER, desarrollamos la programación de varios *Shell Script* que facilitaron enormemente la ardua tarea del preprocesamiento textual. La programación *Shell Script* se eligió por considerarse una programación sencilla, adaptable y reutilizable en múltiples entornos. Estos *scripts* se incorporaron a la aplicación MiNerDoc para automatizar las tareas de preprocesamiento.

Fase Ingeniería de características. Esta fase es la responsable de procesar cada informe clínico a través de la herramienta MetaMap [155]. Distintas técnicas de MT [190] y PLN [1] son realizadas gracias a la herramienta Metamap, entre ellas *part-of-speech tagging*, desambiguación de terminología clínica, interpretación semántica de un texto clínico, resolución de acrónimos, detección de negaciones, etc.

El informe o texto clínico de entrada es fragmentado en oraciones o frases simples y estas a su vez son segmentadas en términos candidatos gracias al servidor tagger de MetaMap (Med-Post / SKR POS) [245]. *Un etiquetador gramatical*, conocido como *POS tagging*, es una herramienta que ayuda a realizar un análisis morfológico que permitirá conocer la categoría gramatical de cada una de las palabras que componen el texto analizado. El servidor tagger de MetaMap (MedPost/SKR POS) es un etiquetador basado en el *tagger* MedPost, desarrollado para dar respuesta al etiquetado de textos biomédicos y entrenado sobre el corpus MEDLINE. MedPost es un etiquetador basado en el modelo oculto de Markov [244]. Gracias a MedPost, el texto original es dividido en oraciones y posteriormente en los diferentes tokens para posteriormente ser etiquetados. Un ejemplo de entrada y salida del etiquetador gramatical de MetaMap puede analizarse en la Figura 3.6. Cada frase del informe clínico proporcionado como entrada, es procesada por el etiquetador de Metamap. Posteriormente, el servidor de desambiguación (WSD) permitirá seleccionar el término candidato más apropiado en

Entrada:

```
This is a test.
```

Salida:

```
[
  ['This', 'det'],
  ['is', 'aux'],
  ['a', 'det'],
  ['test', 'noun'],
  ['.', 'pd']
].
^THE_END^
```

Figura 3.6 .Ejemplo de salida de MedPost/SKR POS de MetaMap

base al contexto desde el que el término fue extraído para poder generar como salida un conjunto de términos con sus variantes, es decir, sinónimos, derivaciones, etc. En base a estos términos y variantes, MetaMap generará un conjunto de conceptos denominados *candidatos* que han sido obtenidos del metatesauro UMLS [195]. Por ejemplo, si introducimos la frase "*pulmonary lymphangitic carcinomatosis*", se obtendrán los siguientes candidatos (ver Figura 3.7):

```
Phrase: pulmonary  lymphangitic carcinomatosis
Meta Candidates (Total=6; Excluded=0; Pruned=0; Remaining=6)

901   Lymphangitic Carcinomatosis (Lymphangitis carcinomatosa)
[Neoplastic Process]
827   Carcinomatosis [Neoplastic Process]
660   pulmonary (Lung) [Body Part, Organ, or Organ Component]
660   Pulmonary (Pulmonary:-:Point in time:^Patient:-) [Clinical
Attribute]
660   Pulmonary (Pulmonary (qualifier value)) [Qualitative Concept]
589   Lymphangitides (Lymphangitis) [Disease or Syndrome]
```

Figura 3.7. Ejemplo de Meta Candidatos

Metamap realiza un ranking de cada candidato denominado "*Meta Mapping*" [263]. Cada candidato obtiene una medida cuyo resultado final va desde el valor 0 hasta el 1000. Esta medida recoge el grado de similitud o exactitud entre una palabra y cada uno de los candidatos contrastados con el metatesauro UMLS. De esta forma un valor de 1000 indicará que la similitud entre el término propuesto y el encontrado por MetaMap es perfecta. Este valor final de cada candidato se obtiene gracias una función de evaluación que ofrece Metamap y que combina cuatro propiedades: *centralidad, variación, cobertura y cohesión.* La *centralidad* determina si los términos analizados forman parte del encabezado de la frase, obteniendo un valor 1 si forman parte del encabezado o 0 si no lo forman. La *variación* es una variable que mide el grado en que difieren las variantes obtenidas de las palabras que constituían la frase original. El valor de la *cobertura* indica que parte del texto de entrada está involucrado en el mapeo final y la *cohesión*, similar a la cobertura, pero enfatiza el cálculo en la

importancia de los componentes que están conectados, teniendo en cuenta los fragmentos de texto contiguo. Por tanto, *MetaMap realiza una interpretación semántica del texto.* Un ejemplo de cómo MetaMap reproduciría la salida del texto "*pulmonary lymphangitic carcinomatosis*", teniendo en cuenta la salida "Meta Mapping" se puede observar en la Figura 3.8.

```
Phrase: pulmonary  lymphangitic carcinomatosis
Meta Mapping (901):
660   Pulmonary (Pulmonary:-:Point in time:^Patient:-)
[Clinical Attribute]
901   Lymphangitic Carcinomatosis (Lymphangitis
carcinomatosa) [Neoplastic Process]
```

Figura 3.8. Ejemplo de Meta Mapping

Basándonos en el concepto de Mapping, hemos realizado dos enfoques para llevar a cabo la tarea de detección de entidades médicas:

- *Detección de todas las posibles entidades médicas:* El primer enfoque selecciona todos los candidatos que MetaMap propone sin seleccionar ningún valor de *mapping*. Con este enfoque conseguiremos mayor expansión de candidatos, y por lo tanto, mayor detección de entidades médicas, aunque puede disminuir la exactitud entre el candidato y el término encontrado en el metatesauro UMLS.
- *Detección de las mejores entidades médicas:* En el segundo enfoque, denominado "Best Mapping", sólo se seleccionaran los candidatos con un mayor nivel de "*Meta Mapping*". Esto significará un mejor ajuste en la calidad de la selección de términos, ya que sólo son seleccionados aquellos candidatos en los que el grado de similitud entre la palabra de la colección y cada uno de los candidatos obtenidos del metatesauro UMLS es totalmente exacto. Dicho con otras palabras, sólo será elegido el candidato que arroje una coincidencia máxima entre el concepto inicial y el que proporciona UMLS.

Fase de extracción de Entidades Médicas y Negaciones. Una de las importantes fuentes de conocimiento que proporciona UMLS es la Red Semántica. En ella cada concepto del metatesauro UMLS se asocia al menos a una categoría semántica de los 133 tipos semánticos existentes [243]. Entre algunas de estas agrupaciones se encuentran las enfermedades, antibióticos, hallazgos, síntomas, etc. Gracias a UMLS, la herramienta MetaMap ofrece, además de la extracción de los candidatos más idóneos desde la información clínica textual, la posibilidad de extraer los tipos semánticos asociados a cada candidato propuesto. En nuestra experimentación sobre la construcción de un sistema MER, hemos seleccionado *17 tipos semánticos* que son los que mejor identifican y centran las distintas entidades médicas consideradas de interés para su posterior aplicación en la detección de factores de riesgo. Por tanto, los 17 tipos semánticos seleccionados y que permiten identifican y extraer los 5 tipos entidades médicas propuestos en esta investigación son los recogidos en la Tabla 3.3.

Tipo Semántico UMLS		Entidad Nombrada
DSYN	Disease or Syndrome	Disease
NEOP	Neoplastic process	Disease
MODB	Mental or Behavioral Dysfunction	Disease
ANTB	Antibiotic	Pharmacologic
PHSU	Pharmacologic Substance	Pharmacologic
BACS	Biologically Active Substance	Pharmacologic
CLND	Clinical Drug	Pharmacologic
BLOR	Body Location or Region	Region/Part Body
BPOC	Body Part, Organ, or Organ Component	Region/Part Body
BSOJ	Body Space or Junction	Region/Part Body
BDSY	Body System	Region/Part Body
DIAP	Diagnostic Procedure	Procedure/Test
LBTR	Laboratory or Test Result	Procedure/Test
LBPR	Laboratory Procedure	Procedure/Test
TOPP	Therapeutic or Preventive Procedure	Procedure/Test
FNDG	Finding	Finding/Sign
SOSY	Sign or Symptom	Finding/Sign

Tabla 3.3. Tipos Semánticos UMLS seleccionados en el sistema MER de MiNerDoc

Al finalizar esta fase se tendrán identificados 5 grupos de entidades médicas: diagnóstico, farmacología, región anatómica, procedimiento/test diagnósticos y síntomas/hallazgos. Adicionalmente, el sistema MER propuesto en esta tesis, permite la *detección de negaciones*. En el dominio médico es de especial relevancia tener presente la detección de negaciones [242]. Es evidente, que no es lo mismo afirmar la presencia de una enfermedad que negarla y por tanto, esta misma lógica es la que hemos plasmado en nuestra aplicación. La detección de la negación supone una de las mejoras más importantes de MetaMap con la incorporación del algoritmo NegEx [241]. Algunos ejemplos de negaciones pueden observarse en el siguiente extracto del informe de alta que se muestra en la Figura 3.9.:

"The patient says that during the course of antibiotics he had decreased hemoptysis. CT on [**2010-09-14**], *showed* ***no evidence of pulmonary embolus*** but did show increased pulmonary nodules with a ground glass appearance. The patient ***denied any chest pain*** but did feel that he had increased pulsations in the neck over the last few days."

Figura 3.9. Ejemplo de negaciones en un informe de alta

En el sistema desarrollado en esta tesis doctoral, hemos detectado estas negaciones y las hemos presentado junto a los cinco grupos de entidades médicas que la herramienta permite extraer, así se reconocerá que entidad médica es realmente una negación según el contenido semántico del informe clínico.

Fase de detección de Factores de Riesgo. En base a las entidades médicas extraídas en el proceso anterior, será posible realizar una detección de los factores de riesgo o alertas clínicas de interés encontradas en cada informe de alta. En el sistema MER desarrollado, el usuario podrá definir inicialmente, en lenguaje natural, los principales factores de riesgo o alertas clínicas asociados a un área concreta de prevención, como por ejemplo, para las enfermedades cardiovasculares se podrían

introducir algunos factores como obesidad, hipercolesterolemia, fumador, etc. Nuestro sistema MER expandirá los términos introducidos, gracias al metatesauro UMLS [195], incorporando sinónimos, variaciones, desambiguación, y otras características que facilitaran la expansión de la búsqueda de factores de riesgo en los informes de alta (ver Figura 3.10). La estrategia de búsqueda se amplía gracias a la expansión de términos que nos permite la herramienta MetaMap. El sistema de MT, una vez detectadas las entidades médicas provenientes del informe de alta original, comparará cada término o entidad médica con los factores de riesgo expandidos, si encuentra coincidencia generará una alerta clínica automática indicado que se ha detectado un factor de riesgo en el informe de alta. Esta generación automática de alertas clínicas apoyará al profesional sanitario en su labor diagnóstica y de prevención temprana. La coincidencia de entidad nombrada y factores de riesgo expandidos se realizará a través de un identificador único de concepto que se encuentra en el metatesauro UMLS denominado CUI (*Concept Unique Identifiers*) [195].

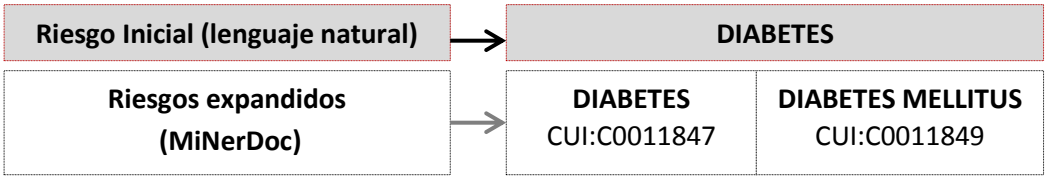


Figura 3.10. Ejemplo de expansión de factores de riesgo iniciales realizado por MiNerDoc

3.3.2. Metodología propuesta para la clasificación diagnóstica multietiqueta.

La metodología propuesta para llevar a cabo la tarea de predicción o asignación automática de códigos de diagnósticos normalizados a informes de alta médica se representa en la Figura 3.11. Para abordar la difícil tarea de la codificación diagnóstica

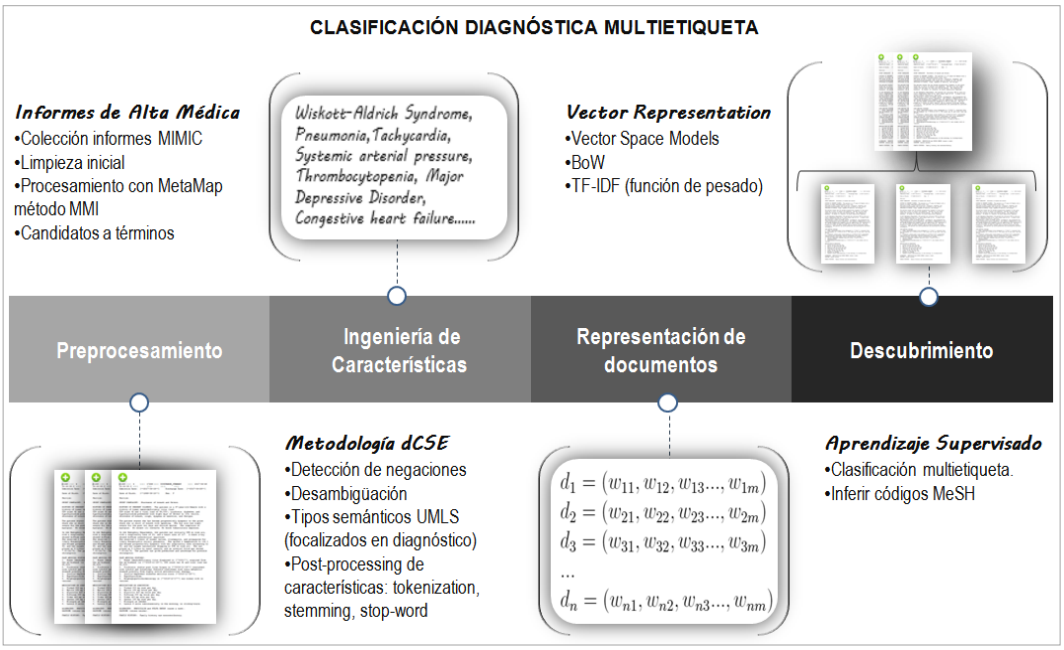


Figura 3.11. Metodología diagnostic Classification with Semantic Enrichement (dCSE)

nos hemos apoyado en diferentes técnicas tomadas de las áreas de la MT [247] y la clasificación multietiqueta [229]. La característica principal de nuestra propuesta se centra en la creación de una nueva metodología que incorpora fuentes externas de conocimiento basadas en metatesauros, como la herramienta MetaMap [155] que se enriquece del metatesauro UMLS [195]. Con el uso de MetaMap y el metatesaurus UMLS conseguiremos generar características de mayor calidad para intentar mejorar notablemente los sistema de categorización convencionales. Estas herramientas externas aportaran ventajas como la resolución de ambigüedades terminológicas, la expansión del vocabulario con sinónimos, la identificación de acrónimos y la detección de negaciones. Nuestra metodología, llamada ***dCSE (diagnostic Classification with Semantic Enrichement)***, sigue cuatro fases básicas: i) *fase de preprocesamiento*, los informes de alta originales son sometidos a una "preparación" previa necesaria para afrontar la fase de procesamiento con MetaMap (eliminación de caracteres no ASCII,

eliminación de líneas en blanco, etc; ii) *fase de ingeniería de características*, se aplican distintas técnicas para obtener características de alta calidad para formar las distintas colecciones de datos; iii) *fase de representación de documentos*, los informes de alta son transformados en un modelo adecuado que permita un análisis eficiente en las siguientes fases de minería; iv) *fase de descubrimiento*, se aplican métodos de aprendizaje multietiqueta (MLL, por sus siglas en inglés) [229,231] para clasificar correctamente los informes de alta médica. Cada una de estas fases serán descritas detalladamente a continuación.

Fase de preprocesamiento. Como hemos comentado anteriormente, la fase de preprocesamiento de una colección textual es una de las fases más laboriosas e importantes en la clasificación automática de documentos. Para llevar a cabo una adecuada selección de rasgos se requiere realizar un proceso previo de transformación de los informes de alta originales. Se procedió a realizar una "preparación" automática del contenido de los 1,210 informes de alta que conforman nuestro dataset, como por ejemplo, eliminación de caracteres no ASCII, supresión de líneas en blanco, y demás características que podrían hacer no funcionar correctamente a la herramienta MetaMap. Posteriormente, la metodología dCSE, aplicará la herramienta MetaMap sobre los informes de alta preprocesados para encontrar conceptos relevantes gracias al metatesauro UMLS. La utilización de la herramienta MetaMap en el proceso de extracción de rasgos de la colección inicial de informes de alta, incorpora un enriquecimiento semántico, aportado por el metatesauro UMLS [195], lo que debe redundar en unos mejores resultados en la categorización de los informes clínicos, este planteamiento será unos de los elementos claves que intentaremos demostrar en la fase de experimentación. Uno de los puntos centrales de la metodología propuesta, a diferencia de la mayoría de los métodos existentes para abordar la tarea de codificación diagnóstica multietiqueta, es el uso del método *MetaMap Indexing* (MMI) [240]. A través de MMI, la aplicación MetaMap procesará la información textual para cada

informe de alta en base a diferentes técnicas de PLN. MetaMap extraerá una serie de términos candidatos desde los informes de alta a través de un proceso complejo que involucra varias fases como tokenización, análisis sintáctico, generación de variantes lingüísticas (sinónimos), evaluación de candidatos y, finalmente, el proceso de *mapping* de conceptos. Después de obtener una lista de términos candidatos, MMI ordenará los resultados de los términos candidatos extraídos a través de una función de ranking [263]. Este valor denominado "*Meta Mapping*" es una medida aplicada a cada candidato y mide el grado de similitud o exactitud entre una palabra y cada uno de los candidatos contrastados con el metatesauro UMLS. Para tener una mayor expansión de términos hemos seleccionado todos los candidatos mapeados a través de la opción MMI. Un ejemplo de salida inicial de un informe de alta de nuestra colección puede observarse en la Figura 3.12. El objetivo final de esta fase es, por lo tanto, obtener un conjunto de términos candidatos de cada informe de alta médica, después de haber sido sometidos a un preprocesamiento inicial para eliminar ruido innecesario que pueda interferir en la tarea de clasificación diagnóstica posterior.

```
00000000|MMI|9.75|Abdomen|C0000726|[blor]|["abd"-tx-19-"ABD"-noun-0]|TX|2324:3|A01.923.047
00000000|MMI|9.75|Heart murmur|C0018808|[fndg]|["murmurs"-tx-19-"murmurs"-noun-
1]|TX|2316:7|C23.888.447
00000000|MMI|9.75|Pneumonia|C0032285|[dsyn]|["Pneumonia"-tx-39-"pneumonia"-noun-
0]|TX|3968:9|C08.381.677;C08.730.610
00000000|MMI|9.74|Nasopharynx|C0027442|[bpoc]|["Nasopharynx"-tx-19-"nasopharynx"-noun-
0]|TX|2158:11|A04.623.557;A14.724.557
00000000|MMI|9.74|Oropharyngeal|C0521367|[blor]|["Oropharynx"-tx-19-"oropharynx"-noun-
0]|TX|2170:10|A04.623.603;A14.724.603
```

Figura 3.12. Salida MMI (MetaMap) de un informe de alta

Fase de ingeniería de características. En esta segunda fase, se analizaran las distintas técnicas llevadas a cabo para realizar la extracción de características o rasgos de mayor calidad procedentes de los informes de alta. Esta es la fase clave que definirá la metodología dCSE. Partiendo de la lista de términos candidatos obtenidos en la fase

anterior, se aplicaran una serie de procedimientos con el objetivo de elegir aquellos rasgos que conformaran el vocabulario final de las distintas colecciones construidas que serán utilizadas en la fase de experimentación para evaluar la tarea CDA propuesta:

- **Detección automática de Negaciones.** Como ya mencionamos en la sección anterior, en el dominio médico es de especial relevancia tener presente la detección de negaciones. MetaMap permite la detección de la negación con la incorporación del algoritmo NegEx [241]. La metodología dCSE incorpora la detección de negaciones en nuestra colección de informes de alta y hemos realizado un proceso automático de eliminación de las mismas de nuestros *datasets* finales ya que, en realidad, la negación de una enfermedad no debería formar parte del proceso de clasificación diagnóstica.

- **Selección de tipos semánticos UMLS.** En esta fase de ingeniería de características es de vital importancia la información facilitada por el metatesauro UMLS y su Red Semántica. Como comentamos en la sección anterior, cada concepto del metatesauro UMLS se asocia al menos a una categoría semántica, dentro de los 133 tipos semánticos existentes, esto supone que cada término se ha categorizado conceptualmente dentro de categorías como síntomas, hallazgos, enfermedades, antibióticos, etc. La herramienta MetaMap permite, a través de UMLS, la posibilidad de extraer los tipos semánticos asociados a cada candidato propuesto. En nuestra experimentación y bajo la metodología propuesta (dCSE), hemos seleccionado *cinco tipos semánticos* (ver Tabla 3.4) que son los que mejor identifican y centran la *categoría diagnóstica*, proporcionando una selección de características de mayor calidad y permitiendo que se genere menos ruido en las colecciones de datos que se utilizaran para llevar a cabo la clasificación diagnóstica. El tipo semántico *DSYN* recoge todos los términos que el metatesauro ha categorizado bajo la etiqueta enfermedad o síndrome, así se obtendría términos como "*diabetes*" o "*myocardial infarction*". El tipo semántico *INPO* incorporará los términos candidatos que UMLS

Tipo Semántico	
DSYN	Disease or syndrome
FNDG	Finding
INPO	Injury or poisoning
NEOP	Neoplastic process
MOBD	Mental or Behavioral Dysfunction

Tabla 3.4. Tipos Semánticos UMLS seleccionados para la tarea de clasificación diagnóstica multietiqueta

ha categorizado bajo la etiqueta de lesiones o envenenamientos causados por agentes externos, pudiendo obtener términos del tipo "*Femoral Fractures*" o "*rib fracture*". El tipo semántico *NEOP* incorpora todos los términos clasificados dentro de los procesos neoplásicos como "*Malignant Neoplasms*" o "*adenocarninoma*". El tipo semántico *MOBD* incorpora los candidatos englobados dentro de la categoría de la disfunción mental o conductual, así bajo esta clase obtendríamos términos como "*Mental disorders*" o "*Psychotic disorders*". Otra característica de la metodología dCSE es que además de incorporar los tipos semánticos más centrados en la categoría diagnóstica, como por ejemplo DSYN (Disease or syndrome) o NEOP (Neoplastic process), se han añadido los rasgos que conforman los hallazgos clínicos descubiertos en los informes de alta incorporando los conceptos que se recogen bajo la categoría semántica *FNDG*.

- **Post-procesamiento de características.** En este procedimiento final, el objetivo es construir un vocabulario adecuado que consiga aumentar el rendimiento de la tarea de clasificación diagnóstica. Se han considerado distintas parametrizaciones, basadas en técnicas de MT, para determinar cuál es la mejor combinación que hace aumentar la eficacia del clasificador (ver Capítulo 5). Las técnicas de MT utilizadas, muy extendidas en la categorización de textos, han sido la eliminación de stop-words, la aplicación de la técnica *stemming* y la utilización de diferentes tipos de tokenizaciones [239]. *La eliminación de stop-words* o eliminación de palabras vacías

hace que desaparezcan de nuestra colección palabras tales como artículos, preposiciones, conjunciones, etc, en definitiva palabras irrelevantes que no aportan calidad a la colección final. Se ha construido una lista propia de stop-words formada por 1,078 términos. *El stemming* es un método de normalización usado para reducir una palabra a su raíz (e.g. *persist* es la raíz de *persisted, persistence* o *persisting*). En nuestra propuesta se ha utilizado uno de los algoritmos más empleados para realizar la técnica de *stemming*, el algoritmo de Porter [110]. A través del proceso de tokenización se descomponen los textos de una colección en unidades mínimas denominadas "*tokens*". En la fase experimental hemos analizado el comportamiento de la tokenización basada en unigramas y bigramas, en la primera el término a extraer es una única palabra y en la basada en bigramas el término a extraer estaría formado por una secuencia de dos palabras. Algunos ejemplos de tokenización basada en unigramas y bigramas presentes en nuestros datasets se recogen en la Tabla 3.5.

Unigramas	Bigramas
infarct	cerebellar infarct
infarcted	infarcted area
infarction	myocardial infarction
infarcts	acute infarcts

Tabla 3.5. Ejemplos de tokenización basada en unigramas y bigramas

La parametrización elegida en la metodología dCSE, y por tanto la aplicada en MiNerDoc, es la creada siguiendo una tokenización basada en unigramas, con eliminación de stop-words y la aplicación de la técnica *stemmin*g [110]. Esta parametrización fue seleccionada entre 8 parametrizaciones distintas al ser la que obtuvo un mejor resultado en la fase experimental que analizaremos en el Capítulo 5.

Fase de representación de documentos. Una vez que todos los informes de alta han sido preprocesados y normalizados, se requiere que se transformen en un modelo adecuado que permita un análisis eficiente en las siguientes fases de minería. La representación es una tarea clave para el procesamiento automático de documentos. Según diferentes investigadores [237,238], una de las representaciones más utilizadas en una amplia mayoría de los trabajos relacionados con la clasificación automática es el modelo de espacio vectorial (VSM). El modelo VSM, propuesto por Salton [126], representa cada documento (d) como un vector simple en un espacio $n - dimensional$ donde cada elemento es una palabra con un peso (w) asociado, $d = (w1, w2, \ldots, wn)$, que representa su importancia dentro de la colección de documentos. La representación más extendida bajo el modelo de espacio vectorial es la conocida *Bag-of-Words* (BoW) [127]. En el sistema CDA de MiNerDoc se ha considerado el modelo *BoW* debido a la buena efectividad demostrada en numerosos estudios [227]. Múltiples funciones de ponderación se han propuesto en la literatura (ver Sección 2.1.3), pero quizás las más usadas sean *Term frequency (TF), Inverse Document Frequency (IDF)* y *Term frequency-Inverse document frequency (TF-IDF)* [128, 129]. En nuestra propuesta hemos elegido la representación vectorial basada en el peso *tf-idf*, en base a los buenos resultados observados en varios estudios centrados en la clasificación de documentación clínica [131, 236].

Fase de descubrimiento. El proceso final de la metodología propuesta (dCSE) es la fase de aprendizaje, donde se logrará inferir la codificación diagnóstica en base al contenido textual de los informes clínicos. El objetivo principal de esta fase es inferir uno o varios códigos de diagnóstico estandarizados (jerarquías diagnósticas MeSH) a partir de las características de un conjunto de datos etiquetados (aprendizaje supervisado). Es importante señalar que la tarea de clasificación diagnóstica se considera principalmente un problema MLL [229] donde cada informe de alta médica puede categorizarse con más de un código de diagnóstico. En esta fase se generará un modelo entrenado por un

conjunto de 1,210 informes de alta de naturaleza multietiqueta provenientes de la colección MIMIC [55]. Este modelo ha sido obtenido aplicando uno de los métodos de transformación de problemas más ampliamente conocidos, el método BR, considerando el algoritmo *Sequential Minimal Optimization* (SMO) como algoritmo de clasificación base. Esta elección se ha respaldado en una extensa fase experimental, que analizaremos en el Capítulo 5 donde se evaluaran un amplio número de modelos multietiqueta teniendo en cuenta los tres métodos existentes en el estado del arte (ver Sección 2.1.4.), métodos de transformación de problemas, métodos de adaptación de algoritmos y métodos basados en multiclasificadores. Además, el modelo elegido (BR con SMO) en la metodología dCSE, ha sido aplicado en otras investigaciones con buenos resultados en la tarea de clasificación multietiqueta [58, 163, 229].

3.4. Funcionalidades

MiNerDoc fue desarrollado con el objetivo de facilitar y apoyar el proceso de toma de decisiones médicas integrando dos objetivos principales:

1. La detección y extracción de conceptos médicos de interés, entidades nombradas, desde el contenido textual de los informes de alta para realizar a través de las entidades extraídas la detección de factores de riesgo o alertas clínicas.
2. La predicción automática de una o varias categorías normalizadas de diagnóstico, descriptores MeSH (*Medical Subject Headings*), que permitirá inferir y automatizar el proceso de codificación diagnóstica multietiqueta.

MiNerDoc está compuesto por dos sistemas, MER y CAD, que integran una gran variedad de funcionalidades que serán analizadas a continuación. En la Figura 3.13 podemos observar la pantalla que da acceso al sistema MiNerDoc a través de la introducción de usuario y contraseña. Una vez introducida la autentificación del usuario

Figura 3.13. Autentificación de usuarios MiNerDoc

aparecerá el interfaz principal de MiNerDoc que dará acceso a todas las funcionalidades que iremos detallando: extracción de entidades médicas, detección de negaciones, detección de alertas o factores de riesgo, clasificación diagnóstica multietiqueta, etc (ver Figura 3.14). Para llevar a cabo cualquiera de las funcionalidades de MiNerDoc, partiremos de la elección del informe (o informes) de alta sobre el que vamos a llevar a cabo las distintas técnicas de MT, NLP y MLL. Para ello, podemos seguir varios puntos de partida:

a) *Seleccionar un informe de alta previamente creado* (ver figura 3.15). MiNerDoc también realiza las funciones básicas de un editor de textos, por lo tanto, simplemente tendremos que ir al icono abrir fichero o seleccionar la opción del menú superior *File* → *Open* y aparecerá una ventana emergente desde la cual seleccionaremos el informe de alta que necesitemos analizar.

b) E*laborar el informe de alta desde el editor de MiNerDoc* (ver Figura 3.16). Si lo que deseamos es crear un nuevo informe de alta en la propia aplicación MiNerDoc lo podemos hacer siguiendo los siguientes pasos: *i)* desde la barra de

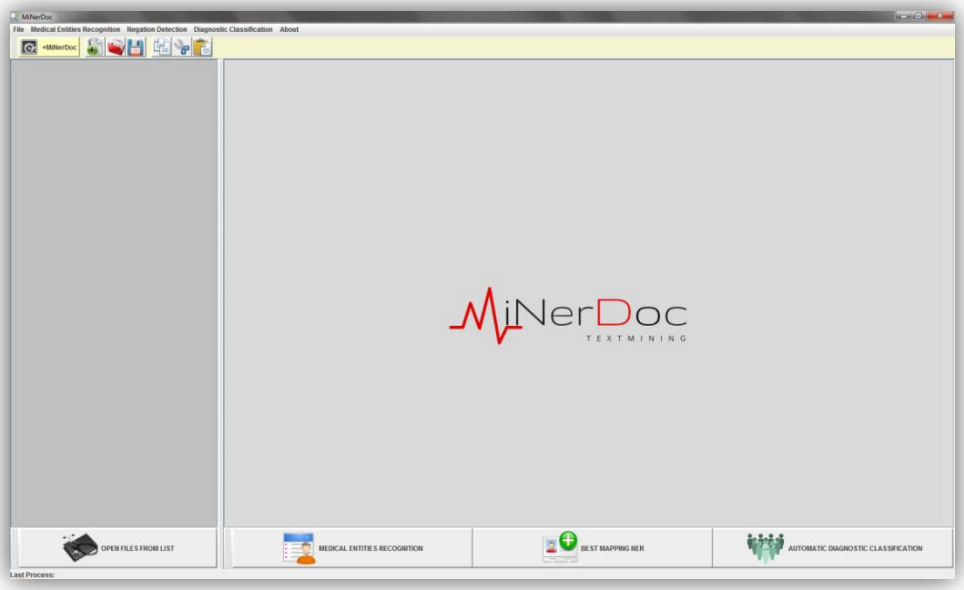


Figura 3.14. Pantalla principal de MiNerDoc. Acceso a todas las funcionalidades.

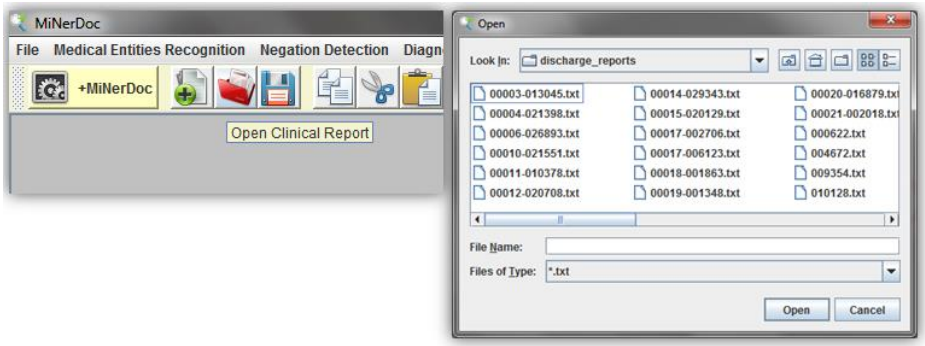


Figura 3.15. Abrir Informe de alta desde editor de MiNerDoc

iconos del menú superior, seleccionar el botón "*Create New Clinical Report*" o *ii*) pulsando en el menú superior sobre la opción "*File→ New*".

c) *Partir de una colección de informes clínicos*. MiNerDoc permite realizar un procesamiento masivo de informes clínicos con la finalidad de realizar una predicción diagnóstica multietiqueta de múltiples informes clínicos con un simple click. Sólo será necesario colocar la colección de informes clínicos en un directorio específico y la aplicación realizará el resto de la tarea de clasificación diagnóstica.

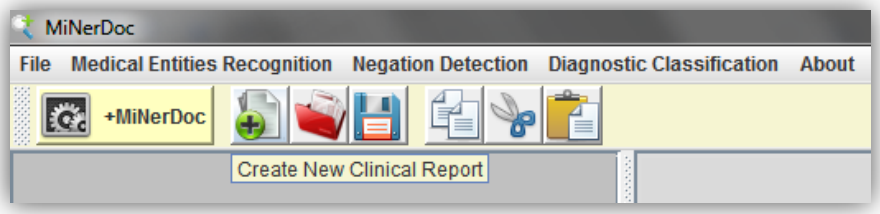


Figura 3.16. Crear nuevo informe de alta

Una vez seleccionado o escrito el informe (o informes) de alta, tal y como hemos comentado en los tres apartados anteriores, podemos realizar las funcionalidades que se recogen en los distintos módulos de la aplicación y que serán descritos a continuación.

3.4.1. Sistema MER de MiNerDoc.

3.4.1.1. Todas las entidades médicas.

Tal y como detallamos en la Sección 3.3 (Metodología), para la detección de las entidades médicas hemos optado por una solución basada en diccionarios gracias a la herramienta MetaMap y el metatesauro UMLS. Hemos implementado dos opciones para realizar la detección de entidades médicas, ambos se diferencian por un concepto muy importante utilizado en MetaMap denominado "*Meta Mapping*" [263]. Como explicamos anteriormente, MetaMap aplica un algoritmo de mapeo que puede ser visto como una puntuación asignada a cada candidato generado para evaluar la importancia de este término con respeto al texto procesado, cuanto mayor es este valor mayor es la concordancia del candidato con el término UMLS. En el módulo de la aplicación MiNerDoc, denominado *"Medical Entities Recognition",* hemos seleccionado todos los candidatos generados por la herramienta MetaMap teniendo en cuenta todos los

términos candidatos mapeados y no sólo los de mayor nivel de concordancia con UMLS. De esta forma, el número de entidades médicas detectadas a través de esta opción será mayor, es decir, habrá una mayor expansión de entidades médicas extraídas (utilizaremos la API de MetaMap para realizar la parametrización correcta).

Hay que destacar que aunque esta expansión de entidades nombradas puede resultar beneficiosa, ya que podemos captar más información de interés para el clínico, puede también llevar asociado una pérdida de exactitud con respecto al término del metatesauro UMLS (el grado de concordancia de la entidad puede no ser totalmente exacto con el término UMLS). Un ejemplo de salida de la aplicación MiNerDoc desde esta opción si introducimos el siguiente fragmento de un informe de alta (Figura 3.17) en el sistema sería el que se observa en la Figura 3.18.

HISTORY OF PRESENT ILLNESS: The patient is a 29 year-old gentleman with a history of depression and multiple suicide attempts who was admitted to the Medical Intensive Care Unit on [**2015-10-10**] for Tegretol and tricyclic antidepressant overdose. The patient has had at least three suicide attempts. He had rheumatoid arthritis. No tobacco use.
The patient was taken to the Operating Room for an exploratory laparotomy and abdominal colectomy and an end ileostomy on [**2015-10-12**] for infarction of the right and transverse colon.

Figura 3.17. Fragmento de informe de alta de la colección MIMIC

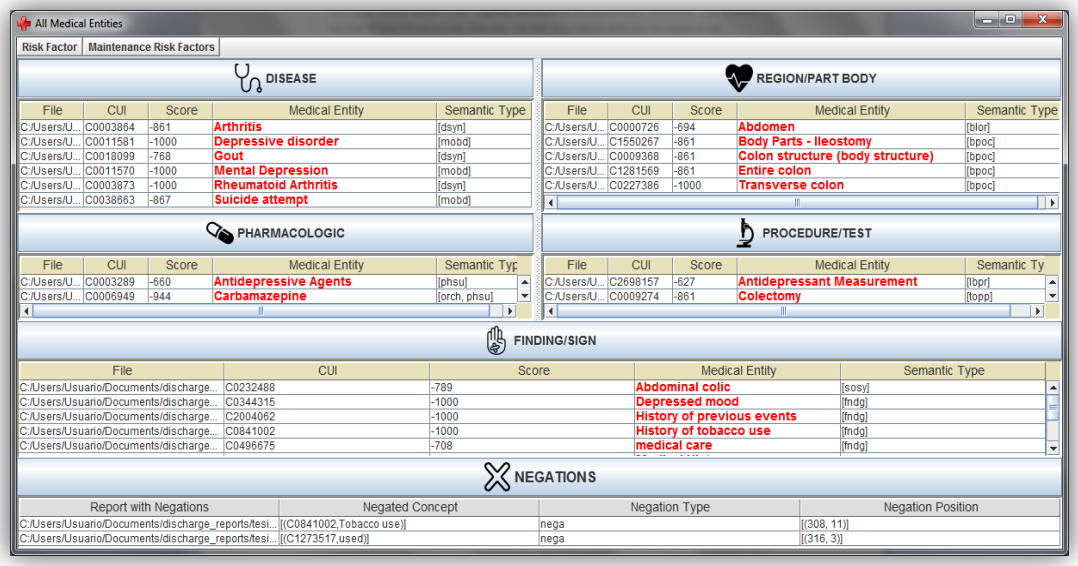


Figura 3.18. Ejemplo salida entidades médicas MiNerDoc, opción todos los candidatos

Como podemos observar, se obtendría un conjunto de entidades médicas para cada tipo de categoría definida en esta investigación: *disease, region/part body, pharmacologic, procedure/test, finding/sign*.

A continuación detallaremos los distintos tipos de entidades médicas que permite extraer esta funcionalidad de MiNerDoc, basándonos en el fragmento de informe recogido en la Figura 3.17. En la Tabla 3.6 podemos observar algunas de las entidades médicas obtenidas del fragmento de informe de alta analizado para las distintas categorías de entidades. Para la entidad "***DISEASE***", MiNerDoc ha obtenido automáticamente las entidades relacionadas con el diagnóstico, como por ejemplo "*Arthritis*" o "*Depressive disorder".* Esta entidad incorpora los tipos semánticos de UMLS denominados *DSYN, NEOP y MODB,* es decir, aquellos que incorporan términos relacionados con enfermedades, síndromes, procesos neoplásicos y enfermedades mentales. Para la entidad "***REGION/PART BODY***", MiNerDoc ha obtenido entre otras los términos "*Abdomen*" o "*Transverse colon"*. Esta entidad incorpora los tipos semánticos

ENTIDAD NOMBRADA "DISEASE"	ENTIDAD NOMBRADA "REGION/PART BODY"
Arthritis	Abdomen
Depressive disorder	Body Parts - Ileostomy
Gout	Colon structure (body structure)
Mental Depression	Entire colon
Rheumatoid Arthritis	Transverse colon
Suicide attempt	
ENTIDAD NOMBRADA "PHARMACOLOGIC"	**ENTIDAD NOMBRADA "PROCEDURE/TEST"**
Antidepressive Agents	Antidepressant Measurement
Carbamazepine	Colectomy
Tegretol	Creation of ileostomy
Tobacco	Exploratory laparotomy
Tricyclic Antidepressive Agents	Laparotomy
ENTIDAD NOMBRADA "FINDING/SIGN"	
Abdominal colic	History of previous events
Depressed mood	History of tobacco use
Deterioration of status	Illness (finding)...

Tabla 3.6. Ejemplo salida entidades médicas, opción MiNerDoc "todos los candidatos"

denominados *BLOR, BPOC, BSOJ y BDSY,* por tanto, incluirían los términos relacionados con la localización o región anatómica, órganos y tejidos. Se recogen además las entidades médicas del grupo "***PHARMACOLOGIC***", algunos términos detectados en el informe analizado y que se encuentran dentro de esta categoría son "*Tegretol*" o "*Antidepressive Agents*". Esta entidad incorpora los tipos semánticos UMLS denominados *ANTB, PHSU, BACS y CLND*, es decir, aquellos términos candidatos que están relacionados con sustancias farmacológicas, antibióticos o sustancias biológicamente activas. En cuanto al grupo "***PROCEDURE/TEST***" se han obtenido automáticamente algunos términos como "*Colectomy*" o "*Laparotomy*". Esta clasificación incorpora los tipos semánticos *DIAP, LBTR, LBPR y TOPP*, que define los procedimientos diagnósticos, procedimientos de laboratorio o procedimientos terapéuticos/preventivos. En la Tabla 3.6 también se observan las entidades médicas de la categoría "***FINDING/SIGN***", en la que se recogen los términos candidatos que tienen relación con los hallazgos, signos y síntomas que han sido encontrado en el informe de alta analizado, como por ejemplo, "*Abdominal colic*" o "*History of tobacco use".*

3.4.1.2. Best Mapping NER.

En el segundo enfoque propuesto, orientado a la extracción de entidades médicas, se seleccionaran los candidatos que tienen las asignaciones de mapeo (mapping) más relevantes (a través de la parametrización de la API de MetaMap). Esto implicará que la calidad y precisión en la selección de entidades médicas será mayor ya que cada candidato extraído tendrá un grado de similitud exacto con el término encontrado en el metatesauro UMLS. Si analizamos la salida de la aplicación MiNerDoc para el mismo fragmento de informe de alta analizado anteriormente (Figura 3.17) podemos observar las entidades médicas detectadas según este enfoque (Figura 3.19). En la Tabla 3.7 se observan los grupos de entidades médicas extraídas en base al fragmento de informe de

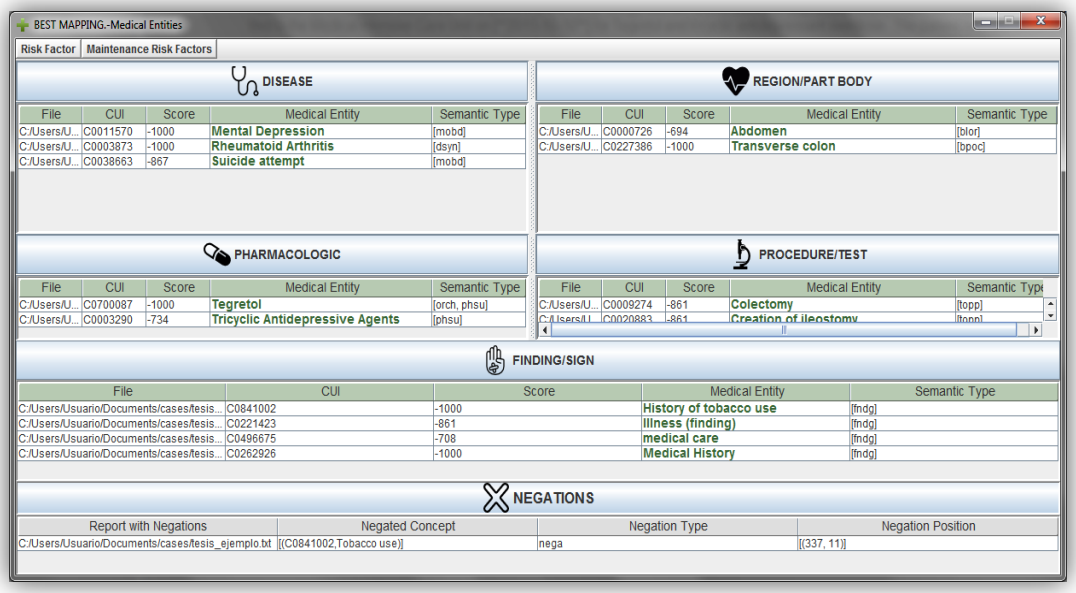


Figura 3.19. Ejemplo salida entidades médicas según opción Best Mapping NER

alta seleccionado tras ser procesado por MiNerDoc. Como podemos ver existe alguna diferencia en la extracción de las entidades médicas entre una opción y otra (todos los candidatos/best mapping). Así, dentro de la categoría "*Disease*", en la opción Best Mapping no se ha detectado el concepto *"Depressive disorder"* o dentro de la categoría "*Region/Part Body*" tampoco se ha detectado el concepto *"Entire colon"*, esto se debe a que no ha existido una concordancia exacta entre el término encontrado en el informe de alta y el concepto UMLS.

ENTIDAD NOMBRADA "DISEASE"	ENTIDAD NOMBRADA "REGION/PART BODY"
Mental Depression	**Abdomen**
Rheumatoid Arthritis	**Transverse colon**
Suicide attempt	
ENTIDAD NOMBRADA "PHARMACOLOGIC"	**ENTIDAD NOMBRADA "PROCEDURE/TEST"**
Tegretol	**Colectomy**
Tricyclic Antidepressive Agents	**Creation of ileostomy**
	Exploratory laparotomy
ENTIDAD NOMBRADA "FINDING/SIGN"	
History of tobacco use	**Medical care**
Illness (finding)	**Medical History**

Tabla 3.7. Ejemplo salida entidades médicas, opción MiNerDoc "Best Mapping NER"

3.4.1.3. Detección de Negaciones.

En el procesamiento de textos clínicos es de especial relevancia la identificación de las negaciones, ya que como es evidente, no es lo mismo afirmar la presencia de una enfermedad que negarla y por tanto, esta misma lógica es la que hemos plasmado en el sistema propuesto. MiNerDoc permite detectar automáticamente estas negaciones a través de dos vías, por un lado, las entidades médicas negadas se podrán visualizar en la misma pantalla (Figura 3.18 y Figura 3.19) donde se muestran los cinco grupos de entidades médicas obtenidas desde los dos módulos de MiNerdoc vistos anteriormente, por otro lado, las negaciones también pueden obtenerse a través de una opción, situada en el menú superior, donde se da la posibilidad de obtenerlas de forma independiente. Para realizar esta funcionalidad, abriremos el informe de alta del cuál necesitemos obtener las negaciones y seleccionaremos desde el menú superior la opción "*Negation detection → Extract Negations from report*" (ver Figura 3.20).

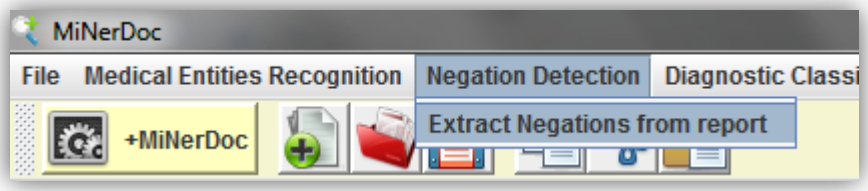


Figura 3.20. Menú Detección de Negaciones

A continuación aparecerá una ventana emergente con los conceptos negados encontrados en el informe seleccionado, en este ejemplo el concepto negado es "*Tobbaco use*" (ver Figura 3.21). Gracias a la herramienta MetaMap es posible detectar estas negaciones para poder ser tratadas adecuadamente. En nuestra aplicación obtenemos los conceptos negados con dos fines, uno de ellos es extraerlos para identificarlos en el proceso de detección de entidades médicas y el segundo objetivo es poderlos excluir del proceso de clasificación diagnóstica. Además de detectar el concepto negado, MiNerDoc presenta otras características como, el identificador CUI, el

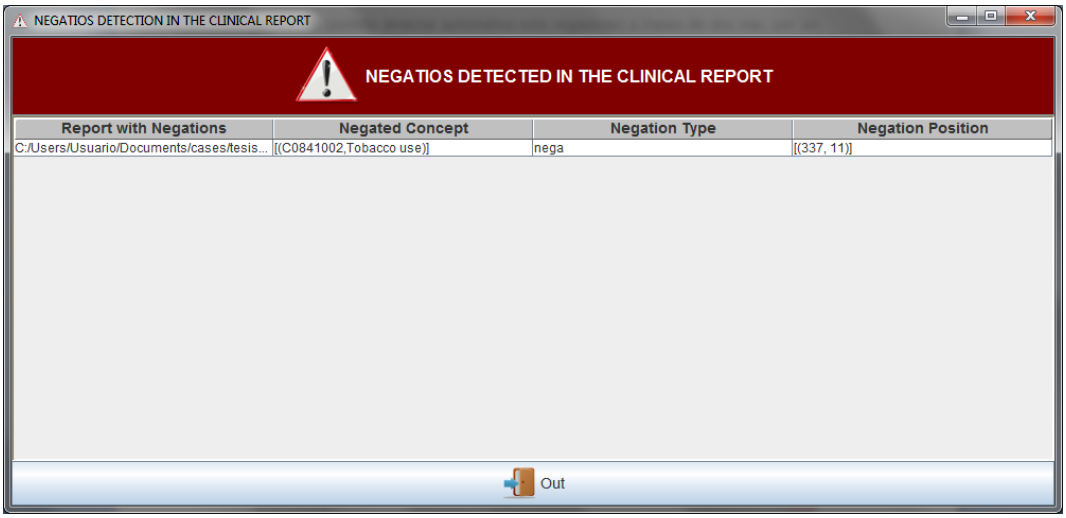


Figura 3.21. Pantalla de detección de negaciones

tipo de negación y la posición del concepto negado, de esta forma, MiNerDoc nos indica donde se encuentra exactamente la negación dentro del informe de alta analizado. Un ejemplo del resultado que obtendríamos al procesar el fragmento de informe de alta recogido en la Figura 3.22 a través de esta funcionalidad de MiNerDoc se muestra en Tabla 3.8.

"The patient says that during the course of antibiotics he had decreased hemoptysis. CT on [**2010-09-14**], *showed* ***no evidence of pulmonary embolus*** but did show increased pulmonary nodules with a ground glass appearance. The patient ***denied any chest pain*** but did feel that he had increased pulsations in the neck over the last few days."

Figura 3.22. Ejemplo de informe de alta con negaciones

DETECCIÓN DE NEGACIONES			
CUI	**CONCEPTO NEGADO**	**POSICION**	**TTPO NEGACIÓN**
C0008031	Chest Pain	(245,10)	nega
C0034035	Pulmonary Embolus	(130,17)	nega

Tabla 3.8. Ejemplo de salida MiNerDoc para detección de negaciones

Tal y como hemos comentado previamente, MetaMap permite la extracción de negaciones gracias a la incorporación del algoritmo *Negex* [241], este algoritmo es capaz de detectar tres tipos diferentes de negaciones [210]:

- ***Términos pseudo-negados (pseudoneg)***. Se tratan de frases que parecen contener términos negados pero que no niegan el término clínico. Si el algoritmo localiza este tipo de negaciones las obvia, saltando al siguiente término negado. Algunos ejemplos de este grupo de negaciones son por ejemplo, *"not only "* o *"not necessarily".* Metamap identifica estas negaciones bajo el tipo *pseudoneg.*

- ***Términos negados PRE-candidatos UMLS.*** Son términos que aparecen antes del concepto UMLS que está negado. Algunos ejemplos de estas negaciones son, *"no sign of", "negative for" o "ruled out".* MetaMap identifica estas negaciones bajo los tipos *nega*, según se trate de un término negado o un término posiblemente negado.

- ***Términos negados POST-candidatos UMLS.*** Se trata de términos que aparecen después del candidato UMLS que está negado. Ejemplos de este tipo de negaciones son, *"unlikely" o "free".* MetaMap identifica estas negaciones bajo los tipos *negb*, según se trate de un término negado o un término posiblemente negado.

Como podemos observar en la Tabla 3.8 se han detectado dos negaciones del tipo "nega" (las más habituales), este tipo de negaciones son las que MetaMap clasifica, en base al algoritmo Negex, cómo términos que hacen negativa la frase y aparecen previamente al concepto UMLS negado, por ejemplo el caso recogido en el ejemplo "*"no evidence of pulmonary embolus"* es un tipo de negación *"nega".* Otros ejemplos de distintos tipos de negaciones pueden observarse en la Tabla 3.9.

TIPOS DE NEGACIONES	
TEXTO ANALIZADO	**TTPO NEGACIÓN**
Cyanosis free	negb
No evidence of cianosis	nega
Not only cianosis	pseudoneg

Tabla 3.9. Ejemplos de tipos de negaciones MetaMap

3.4.1.4. Detección de factores de riesgo.

En base a las entidades médicas extraídas (diagnóstico, farmacología, procedimientos, síntomas y localización anatómica), MiNerDoc puede detectar automáticamente factores de riesgo o alertas clínicas de interés de distintos ámbitos asistenciales tomando como base el contenido textual de cada informe clínico. Esta funcionalidad se encuentra disponible dentro de los dos módulos descritos anteriormente (opción todos los candidatos y opción Best Mapping NER). A continuación, analizaremos los pasos o procesos necesarios para llevar a cabo la detección de factores de riesgos a través de los siguientes módulos de MiNerDoc, mantenimiento de factores de riesgo iniciales y detección de factores de riesgo:

Mantenimiento de factores de riesgo de partida. Inicialmente el usuario puede definir en lenguaje natural los principales factores de riesgo o alertas clínicas asociadas a un área concreta de prevención, como por ejemplo, las enfermedades del corazón o las enfermedades respiratorias. Esta acción únicamente será necesaria realizarla la primera vez que utilicemos la herramienta o cuando queramos ampliar los factores de riesgo que se definieron inicialmente. Para llevar a cabo esta función, el usuario, a través del módulo de mantenimiento de factores de riesgo de MiNerDoc, puede dar de alta los factores de riesgo iniciales y visualizar posteriormente estos factores y los términos expandidos gracias a MetaMap. Los módulos para llevar a cabo estas tareas se detallan a continuación:

- ***Alta Factores Riesgo iniciales.*** El usuario, en su primera toma de contacto con la aplicación, dará de alta los factores de riesgo más habituales asociados con las enfermedades cardiovasculares o las enfermedades respiratorias, introduciendo términos como por ejemplo, obesidad, hipercolesterolemia, fumador, etc (ver Figura 3.23). Una vez dado de alta los factores de riesgos iniciales, la aplicación MiNerDoc interpretará este término, descrito en lenguaje natural, y buscará el término más adecuado según el metatesauro UMLS, expandiendo estos términos iniciales (con todas las derivaciones lingüísticas) y asignándolo un identificador único de concepto denominado CUI (*Concept Unique Identifier*). Para tener una mayor cobertura en la detección de riesgos se configuró la API de MetaMap para extraer todas las derivaciones lingüísticas del término introducido inicialmente por el usuario (además de añadir la detección de acrónimos y opción de desambiguación).

La *expansión* automática de los factores de riesgo iniciales (introducidos en lenguaje natural), gracias a la API de MetaMap, nos permitirá encontrar un mayor número de alertas clínicas en cada informe de alta.

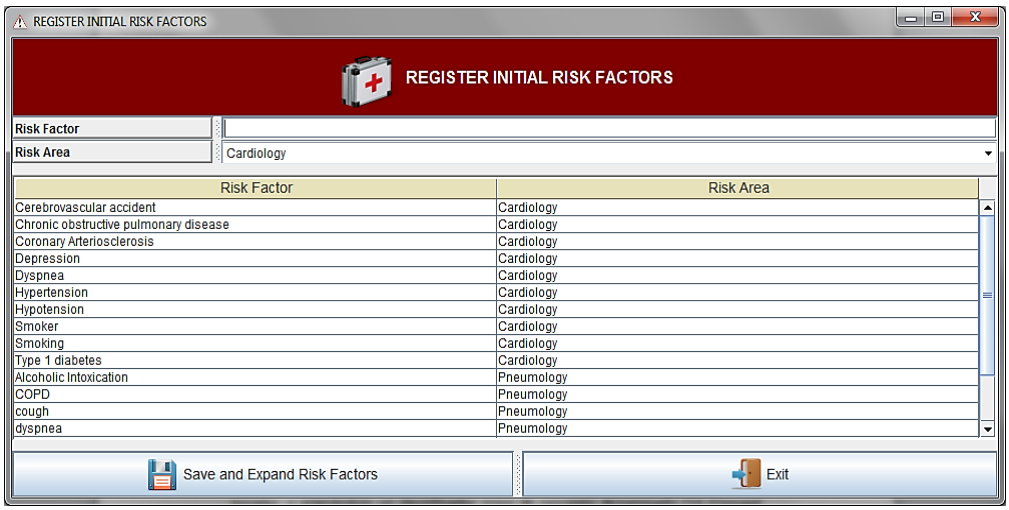


Figura 3.23. Alta de Factores de Riesgo Iniciales (Maintenance risk factors)

La *expansión* automática de los factores de riesgo iniciales introducidos en lenguaje natural, gracias a la api de MetaMap, nos permitirá encontrar un mayor número de alertas clínicas en cada informe de alta. Así por ejemplo, si se introduce la palabra "*diabetes*" en la opción "*Maintenance Risk Factor → Add Initial Risk Factor*", tanto en el módulo "*Medical Entities Recognition*" como el módulo "*Best Mapping NER*", automáticamente se realizará una expansión de términos en base al metatesauro UMLS y añadiéndose a la estrategia de búsqueda los términos *"diabetes", "diabetic"* y *"diabetes mellitus".* En la Figura 3.24 se recoge un ejemplo de expansión de términos si introducimos el texto "*depression*" en esta funcionalidad de MinNerdoc.

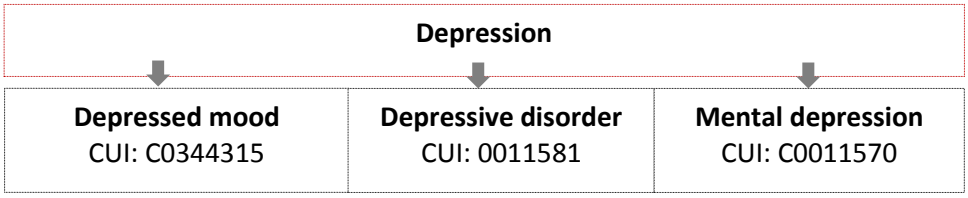


Figura 3.24. Ejemplo de expansión de términos

- ***Visualizar factores de riesgo expandidos.*** Otra de las opciones que puede realizarse en relación al mantenimiento de factores de riesgo es la visualización de los términos una vez realizado el proceso de expansión de los factores de riesgos iniciales. Para ello seleccionaremos, una vez llevado a cabo la detección de entidades médicas, la opción "*Maintenance Risk Factor → Visualize Expanded Risk Factor*" y se visualizaran todos los términos expandidos, gracias a MetaMap y UMLS, para realizar la detección final de los factores de riesgo encontrados en el informe de alta analizado. En la Figura 3.25 podemos observar un ejemplo de visualización de los factores de riesgos expandidos con MetaMap, donde se recoge el término CUI, el nombre del término expandido y el área de riesgo en el que se ha definido.

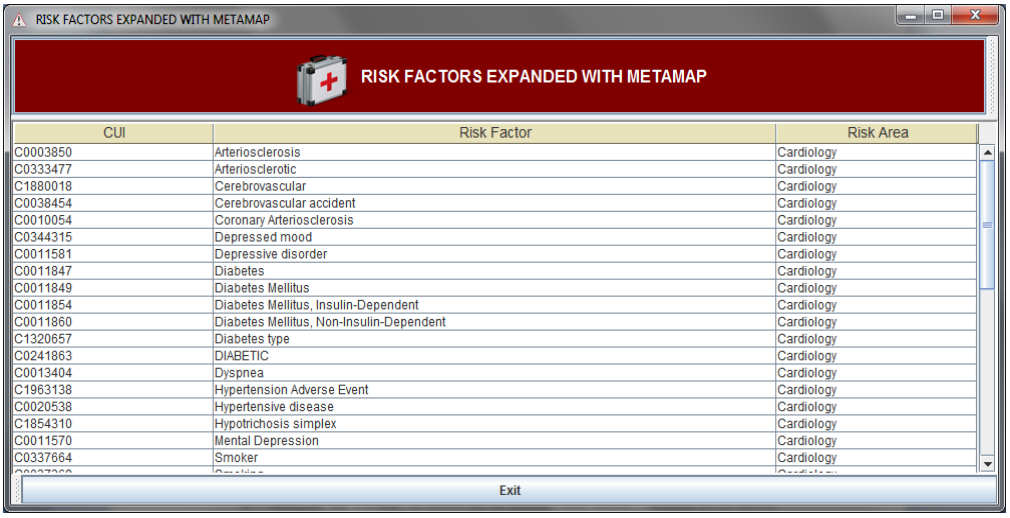
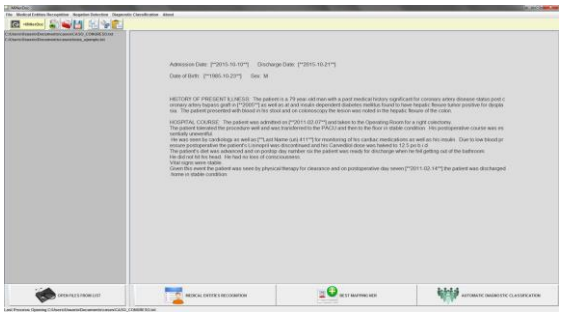


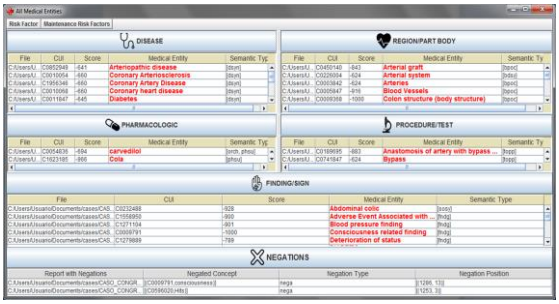
Figura 3.25. Visualizar Factores de Riesgo expandidos (Maintenance Risk Factor)

Detección de Factores de Riesgo (enfermedades del corazón y enfermedades respiratorias). Una vez definidos los factores de riesgo iniciales, MiNerDoc podrá detectar automáticamente las alertas o factores de riesgo que puedan encontrarse en cualquiera de los informes de alta analizados, todo ello en base a las distintas entidades médicas extraídas de dichos informes. El sistema comparará las entidades médicas encontradas en el informe de alta con los factores de riesgo expandidos por MetaMap, si existe coincidencia entre ambos términos según el identificador de UMLS (CUI) el sistema generará una alerta clínica automática. La generación automática de estas alertas clínicas facilitará la tarea de extracción de conocimiento en base al contenido textual de los informes clínicos y facilitará la toma de decisiones clínicas, todo ello con la ventaja de reducir tiempo y evitar posibles errores humanos.

En la aplicación MiNerDoc hemos focalizado la búsqueda de factores de riesgo en base a dos categorías, enfermedades del corazón y enfermedades respiratorias, aunque estas categorías podrían ampliarse fácilmente en futuras mejoras. En la figura 3.26 se recoge el procedimiento para detectar automáticamente los factores de riesgo de un informe.

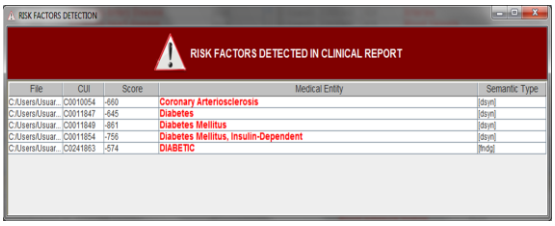


Paso 1. Introducir en MiNerDoc el informe de alta que se desea analizar para extraer las entidades médicas y para detectar los factores de riesgo que puedan estar presentes en el informe.



Paso 2. Entrar en alguna de las opciones que permiten obtener la entidades médicas ("*Medical Entities Recognition*" o "*Best Mapping NER*").

Paso 3. Ejecutar opción "*Risk Factor*" y seleccionar el ámbito donde deseemos buscar las alertas clínicas (enfermedades cardiacas o respiratorias). Previamente tienen que estar definidos los factores iniciales de alerta por el usuario.



Paso 4. Si se detectan factores de riesgo, MiNerDoc mostrará una pantalla con la lista de los factores de riesgos encontrados en el informe.

Figura 3.26. Esquema del proceso a seguir para la obtención de factores de riesgo

De este modo, si procesamos el siguiente fragmento de informe de alta (Figura 3.27),

Admission Date: [**2015-10-10**] Discharge Date: [**2015-10-21**]

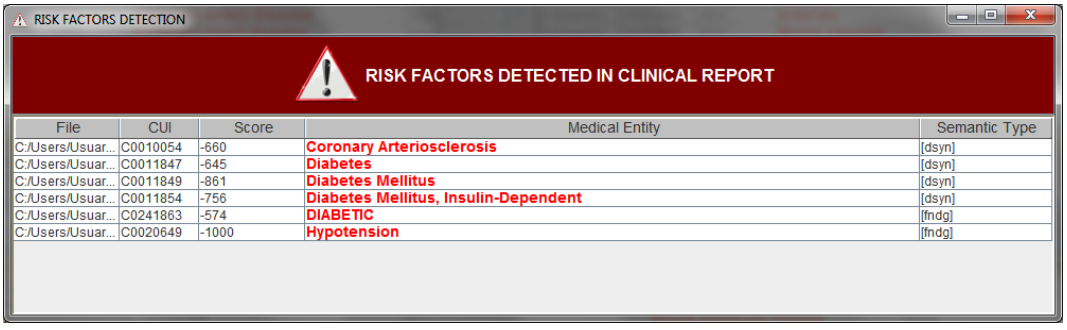
Date of Birth: [**1985-10-23**] Sex: M

HISTORY OF PRESENT ILLNESS: The patient is a 79 year-old man with a past medical history significant for coronary artery disease status post coronary artery bypass graft in [**2005**] as well as at and insulin dependent diabetes mellitus found to have hepatic flexure tumor positive for dysplasia. The patient presented with blood in his stool and on colonoscopy the lesion was noted in the hepatic flexure of the colon.

HOSPITAL COURSE: The patient was admitted on [**2011-02-07**] and taken to the Operating Room for a right colectomy.
The patient tolerated the procedure well and was transferred to the PACU and then to the floor in stable condition. His postoperative course was essentially uneventful.
He was seen by cardiology as well as [**Last Name (un) 411**] for monitoring of his cardiac medications as well as his insulin. Due to low blood pressure postoperative the patient's Lisinopril was discontinued and his Carvedilol dose was halved to 12.5 po b.i.d.
The patient's diet was advanced and on postop day number six the patient was ready for discharge when he fell getting out of the bathroom. He did not hit his head. He had no loss of consciousness. Vital signs were stable. Given this event the patient was seen by physical therapy for clearance and on postoperative day seven [**2011-02-14**] the patient was discharged home in stable condition.

Figura 3.27. Ejemplo de Informe de alta de la colección MIMIC

obtendríamos automáticamente los siguientes factores de riesgo o alertas relacionadas con las "enfermedades del corazón" (ver Figura 3.28), por ejemplo "*Coronary Arteriosclerosis*", "*Hypotension*", y ninguna alerta relacionada con las "enfermedades respiratorias" (en base a los factores de riesgo iniciales introducidos en la aplicación).



File	CUI	Score	Medical Entity	Semantic Type
C:/Users/Usuar...	C0010054	-660	Coronary Arteriosclerosis	[dsyn]
C:/Users/Usuar...	C0011847	-645	Diabetes	[dsyn]
C:/Users/Usuar...	C0011849	-861	Diabetes Mellitus	[dsyn]
C:/Users/Usuar...	C0011854	-756	Diabetes Mellitus, Insulin-Dependent	[dsyn]
C:/Users/Usuar...	C0241863	-574	DIABETIC	[fndg]
C:/Users/Usuar...	C0020649	-1000	Hypotension	[fndg]

Figura 3.28. Factores de riesgo detectados en el informe analizado por MiNerDoc

3.4.2. Funcionalidades del sistema CDA

En este trabajo hemos intentado acercar al área de la Medicina computacional una de las tareas más ampliamente utilizadas en el ámbito de la minería de textos como es la clasificación automática de documentos. Por este motivo, hemos desarrollado una funcionalidad, dentro de nuestro sistema MiNerDoc, que permite inferir uno o más códigos de diagnóstico estandarizados (descriptores MeSH asociados a enfermedades) a partir de las características textuales de uno o múltiples informes clínicos. Para llevar a cabo esta tarea hemos seguido un enfoque basado en el paradigma del aprendizaje multietiqueta, añadiendo la peculiaridad de incorporar recursos externas de conocimiento a través de la herramienta MetaMap y el metatesauro UMLS.

Esta funcionalidad de MiNerDoc se basa en la metodología propuesta en esta tesis doctoral, denominada dCSE (*diagnostic Classification with Semantic Enrichment*), y que ha sido explicada anteriormente en la Sección 3.3.2. La metodología dCSE será evaluada en el Capítulo 5 para determinar el rendimiento predictivo del modelo de clasificación desarrollado en esta tesis.

La clasificación diagnóstica desde MiNerDoc se puede realizar de dos formas distintas, una en la que partiremos del contenido textual de un informe de alta y otra en la que será posible realizar una clasificación diagnóstica masiva al considerar un conjunto de informes clínicos. A continuación, detallaremos los módulos de MiNerDoc desarrollados para llevar a cabo ambos enfoques.

3.4.2.1. Informe clínico único.

A través de este módulo, MiNerDoc realiza la asignación automática de códigos de diagnósticos normalizados (22 descriptores de enfermedad MeSH) en base al contenido textual de un informe clínico. Los pasos que el usuario deberá seguir para realizar esta tarea serán detallados a continuación:

a) Seleccionar un informe de alta previamente creado o elaborar uno desde el editor de informes de MiNerDoc (Figura 3.29).

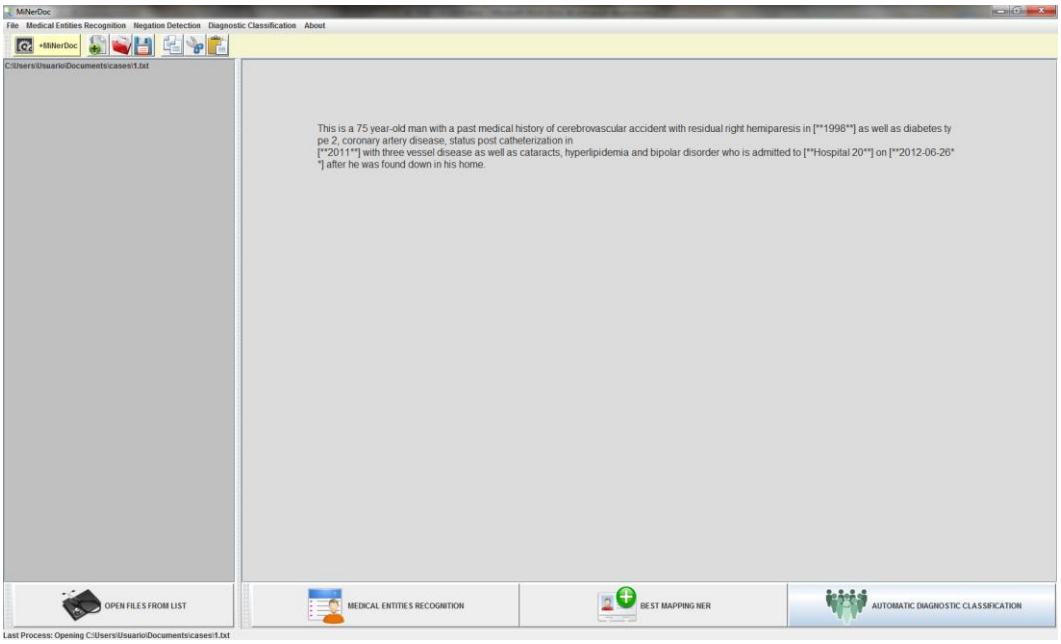


Figura 3.29. Abrir informe de alta para realizar clasificación diagnóstica automática

b) Seleccionar opción del menú principal "Automatic Diagnostic Classification". Una vez seleccionado el informe de alta sobre el que deseemos obtener la predicción diagnóstica, pulsaremos sobre el botón inferior del menú principal *"Automatic Diagnostic Classification*" (ver Figura 3.31). Esta opción nos permitirá llevar a la práctica el desarrollo de la metodología propuesta en nuestra tesis, metodología dCSE. Una vez seleccionada esta opción aparecerá una ventana emergente (ver Figura 3.30) donde se podrá optar por dos vías: i*) ejecutar el proceso de clasificación paso a paso*, podemos realizar el proceso predictivo siguiendo cada una de las tareas de MT necesarias para llegar a la predicción final (ver Sección 3.3.2); ii) o ejecutando el *proceso completo* de predicción, donde con un click realizaremos todas los tareas de MT en un solo paso.

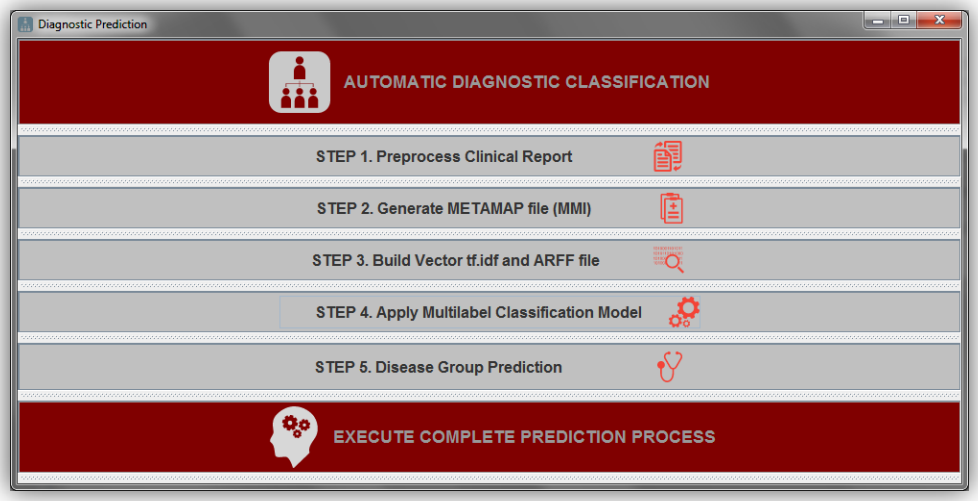


Figura 3.30. Pantalla que inicia el proceso de predicción diagnóstica multietiqueta.

Ambas opciones, tanto la vía paso a paso o la ejecución completa, realizarán las siguientes tareas:

- *Preprocesamiento inicial del informe clínico.* En esta fase MiNerDoc realiza una "limpieza" automática del contenido de los informes de alta, realizándose tareas como, la eliminación de caracteres no ASCII, supresión de líneas en blanco, y la eliminación de otras características que podrían hacer no funcionar correctamente a la herramienta MetaMap.

- Generación automática del fichero de conceptos centrados en el diagnóstico. Basándonos en la utilidad proporcionada por MetaMap denominada MMI (MetaMap indexing) [240], seremos capaces de extraer del informe de alta original una serie de términos candidatos, centrados en el diagnóstico, obtenidos gracias a la selección de aquellos candidatos asociados a cinco tipos semánticos UMLS (ver Sección 3.3.2).

- *Fase de ingeniería de características.* En esta fase se llevan a cabo una secuencia de tareas comunes en MT, como la detección de negaciones, la eliminación de stop-words, las diferentes formas de tokenización, la aplicación de la técnica *stemming*, etc. MiNerDoc está configurada para llevar a cabo una ingeniería de características con una parametrización que sigue una tokenización basada en unigramas, con eliminación de *stop-words* y aplicación de la técnica *stemming*. Esta elección se basó en los buenos resultados obtenidos en la fase experimental que veremos en el Capítulo 5, donde comprobaremos como dicha parametrización mejoró significativamente el rendimiento de la clasificación diagnóstica multietiqueta.

- *Representación de documentos.* MiNerDoc lleva a cabo la representación de documentos aplicando el modelo *BoW* (añadiendo la fase previa de enriquecimiento semántico) y la representación vectorial basada en la función de ponderación tf-idf, ambos elegidos por ser considerados en numerosos estudios sobre categorización de textos [227].

- *Fase de descubrimiento.* En esta fase, MiNerDoc aplica el modelo de clasificación multietiqueta elegido para alcanzar la predicción diagnóstica final del informe de alta elegido. Para ello, el sistema deberá inferir uno o varios códigos de diagnóstico estandarizados a partir de las características obtenidas de dicho informe clínico y del conjunto de características de entrenamiento proveniente de 1,210 informes clínicos de la colección MIMIC [55] previamente etiquetados (aprendizaje supervisado). En nuestro caso, para llevar a la práctica la tarea de clasificación multietiqueta hemos seleccionado el método de transformación de problemas BR utilizando el algoritmo *Sequential Minimal Optimization* (*SMO)* como algoritmo de clasificación base. Esta elección se fundamenta en varias

premisas, una de ellas es que los métodos de transformación de problemas, y en concreto el método BR, ofrece un rendimiento óptimo en escenarios clínicos específicos [57] y es uno de los métodos más utilizados en la resolución de problemas de clasificación multietiqueta [58]. Del mismo modo, la utilización del algoritmo base SVM (SMO) garantiza buenos resultados bajo aquellos *datasets* que tienen un alto número de características y un pequeño número de instancias (como es nuestro caso) [229]. Además de lo citado anteriormente, la elección (BR+SMO) se respalda en el extenso análisis experimental que será descrito en el Capítulo 5.

Una vez seleccionada la opción deseada para comenzar el proceso de predicción diagnóstica, aparecerá un mensaje que determinará si el informe clínico ha sido clasificado o si por el contrario no ha podido ser clasificado por no encontrar ninguna categoría diagnóstica apropiada. A continuación, se mostrará una nueva pantalla donde aparecerán, destacados en rojo, los distintos grupos de diagnóstico MeSH en los que se ha clasificado el informe (ver Figura 3.31).

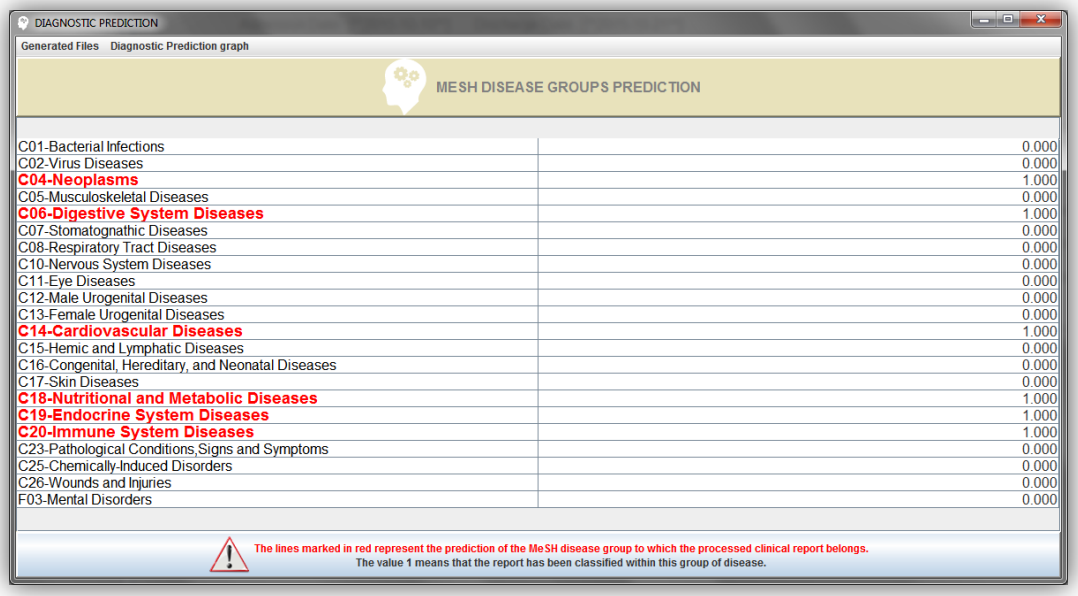


Figura 3.31. Pantalla que muestra el resultado final de la predicción diagnóstica multietiqueta

En este ejemplo, MiNerDoc ha clasificado el informe de alta en seis grupos de diagnósticos MeSH: *C04-Neoplasms, C06-Digestive System Diseases, C14-Cardiovascular Diseases, C18-Nutritional and Metabolic Diseases, C19-Endocrine System Diseases* y *C20-Immune System Diseases*. Una vez obtenida la predicción diagnóstica final, MiNerDoc permite realizar dos nuevas funcionalidades, visualizar los ficheros test y training utilizados para llevar a cabo la clasificación diagnóstica y visualizar una representación gráfica en relación al proceso de clasificación. Ambas funcionalidades las detallaremos a continuación.

c) *Fase final de la predicción diagnóstica. Ficheros test/training generados y representación gráfica de la clasificación.* Una vez obtenida la lista de las categorías diagnósticas en las que se ha clasificado el informe de alta seleccionado nos encontramos en el menú superior con dos nuevas funcionalidades: "*Generated Files*" y "*Diagnostic Prediction graph*" (ver Figura 3.32).

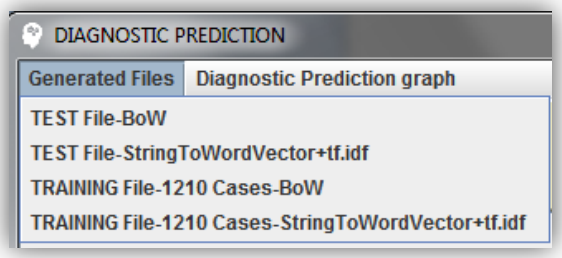


Figura 3.32. Menú superior Proceso Clasificación diagnóstica Multietiqueta. Opciones "Generated Files" y "Diagnostic prediction graph".

- *Ficheros Generados (Generated Files).* MiNerDoc realiza la construcción automática de los ficheros test y training (extensión arff) que son utilizados para la construcción de la predicción diagnóstica. De esta forma a través de esta opción, podrán ser visualizados los distintos ficheros que han sido utilizados en el proceso de clasificación, como el fichero test de bolsa de palabras (*TEST file BoW*) que es construido gracias a la conceptualización diagnóstica del informe

clínico procesado por MetaMap. Un ejemplo de fichero test generado por la aplicación MiNerDoc, en base al informe de alta de la Figura 3.27 se representa en la Figura 3.35. Además será posible visualizar los ficheros generados después de realizar la fase de representación de documentos basándonos en el peso tf-idf. Estos ficheros, si se desean, podrán ser utilizados directamente en las aplicaciones Meka y Mulan (herramientas que permiten abordar el aprendizaje multietiqueta) para llevar a cabo otros análisis externos.

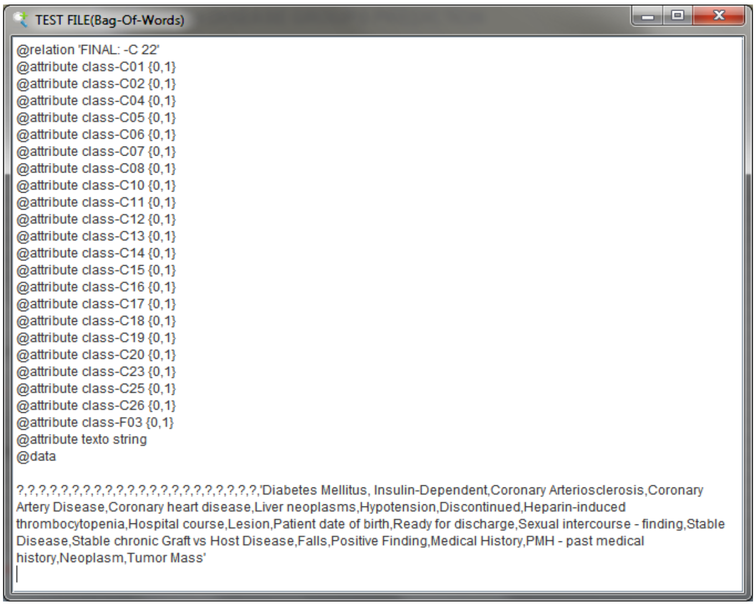


Figura 3.33. Menú superior Proceso Clasificación diagnóstica Multietiqueta. Opción Generated Files->TEST file (BoW)

- *Gráfico de la predicción diagnóstica (Diagnostic Prediction graph).* La aplicación MiNerDoc nos permite visualizar gráficamente el resultado de la clasificación diagnóstica basándose en la representación del top ten de los términos candidatos que han sido seleccionados automáticamente para llevar a cabo la predicción diagnóstica (Figura 3.34). Para llevar a cabo esta representación, se

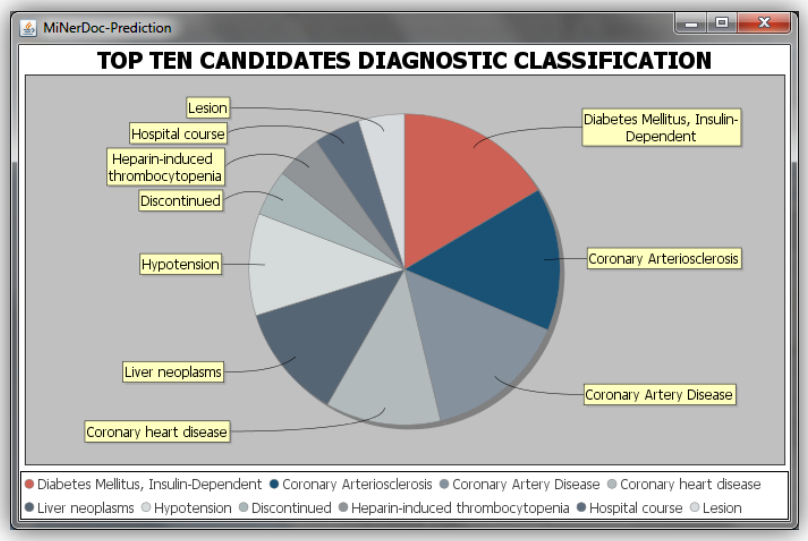


Figura 3.34. Gráfico Candidatos Clasificación (Top ten candidatos).

realiza una ordenación de los términos candidatos que han formado parte de la clasificación en base a la puntuación (*score*) que ha asignado MetaMap según la funcionalidad MMI. Las 10 mejores puntuaciones de esta función de ranking que proporciona MMI serán las que se representen en el gráfico. Estos valores representados (los más altos del ranking) han sido los que han tenido mayor grado de similitud o exactitud con el término del metatesauro UMLS. Por tanto, este gráfico nos puede dar una visión global y fiable de los conceptos relacionados con los diagnósticos más destacados de un informe clínico. Hay que recordar que MiNerDoc utiliza 5 tipos semánticos que son los que mejor identifican y centran la categoría diagnostica, por tanto, estos 5 tipos semánticos también son los que se plasman en esta representación gráfica. Tal y como ya se describió en la Sección 3.3.2, los tipos semánticos utilizados en todo el proceso de clasificación propuesto son DSYN (*disease or syndrome*), INPO (I*njury or poisoning*), NEOP (*Neoplastic process*), FNDG (*Finding*) y MODB (*Mental or B. Dysfunction*).

3.4.2.2. Múltiples informes clínicos

Como hemos comentado anteriormente, una de las funcionalidades más destacadas de MiNerDoc es la de poder realizar automáticamente un proceso de categorización diagnóstica en base a la información de contenido textual procedente de un informe clínico. Para realizar esta tarea nuestra aplicación se sirve de la herramienta MetaMap y del metatesauro UMLS, fuentes de conocimiento que aportan un enriquecimiento semántico que hace mejorar el rendimiento predictivo. Hemos desarrollado una extensión de esta funcionalidad para poder realizar un procesamiento masivo con la finalidad de realizar una predicción diagnóstica de un conjunto de informes clínicos. Para llevar a cabo esta funcionalidad hemos seguido los pasos de la metodología propuesta en esta tesis doctoral, denominada dCSE, y que ha sido anteriormente explicada en la Sección 3.3.2. A continuación, detallaremos las principales características de esta funcionalidad a la que hemos denominado "*+MiNerDoc*". Los pasos para conseguir realizar esta predicción diagnóstica masiva son los siguientes:

a) Seleccionar opción del menú principal +MiNerDoc. Está opción se sitúa tanto en el menú superior de la pantalla principal como en la barra de herramientas. Una vez pulsada esta opción, aparecerá una ventana emergente con los pasos que tendrá que seguir el usuario final para realizar la predicción diagnóstica de múltiples informes clínicos (ver Figura 3.35). Lo único que MiNerDoc necesita es que todos los informes clínicos que deseemos procesar estén ubicados en el directorio indicado, es necesario que los informes tengan la extensión .txt. Desde la pantalla anterior podemos visualizar en todo momento los informes clínicos que han sido seleccionados para realizar el proceso masivo de clasificación, para ello pulsaremos sobre la opción del menú superior "*Report +MiNerDoc*" (ver Figura 3.36). Los informes podrán ser también editados desde esta opción.

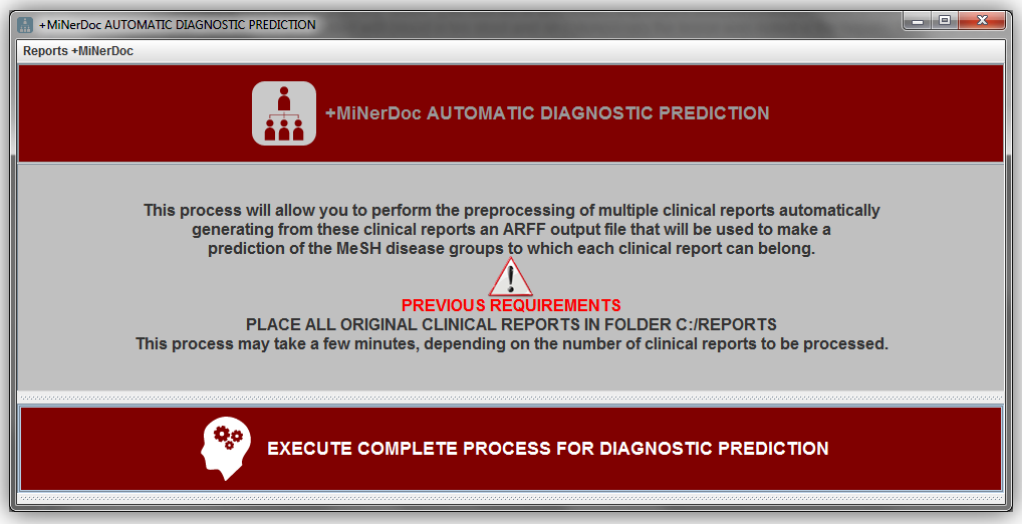


Figura 3.35. Ventana que inicia el proceso de clasificación diagnóstica masiva, opción +MiNerDoc.

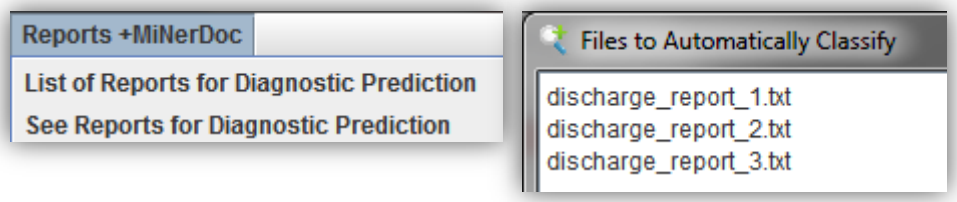


Figura 3.36. Menú superior de la Opción "+MiNerDoc".

b) Seleccionar opción "Ejecutar proceso completo predicción diagnóstica (execute complete process for diagnostic prediction)". Una vez seleccionados y ubicados en el directorio correspondiente los informes clínicos sobre los que queramos realizar la predicción diagnóstica, pulsaremos sobre el botón inferior de la pantalla inicial (Figura 3.35) y se realizaran los siguientes procesos de MT para cada informe clínico:

1. Preprocesamiento inicial de todos los informes clínicos.
2. Generación automática del fichero de conceptos centrados en el diagnóstico para cada informe clínico (salida MMI de MetaMap).

3. Fase de ingeniería de características de cada informe clínico.
4. Representación vectorial de los informes clínicos procesados.
5. Fase de descubrimiento de las categorías diagnosticas normalizadas para cada informe.

Una vez finalizado el proceso aparecerá una pantalla donde se visualizará la predicción diagnóstica de la colección de informes clínicos, mostrándose una línea con la predicción de cada informe de alta (ver Figura 3.37). Como podemos observar, aparecerá una línea por cada uno de los informes clínicos clasificados, mostrándose el nombre de cada informe textual procesado y marcando con 1 las categorías MeSH en las que se ha clasificado dicho informe. En este ejemplo, se mostraría el resultado de tres informes clínicos que han sido clasificados en las siguientes jerarquías MeSH (ver Tabla 3.10):

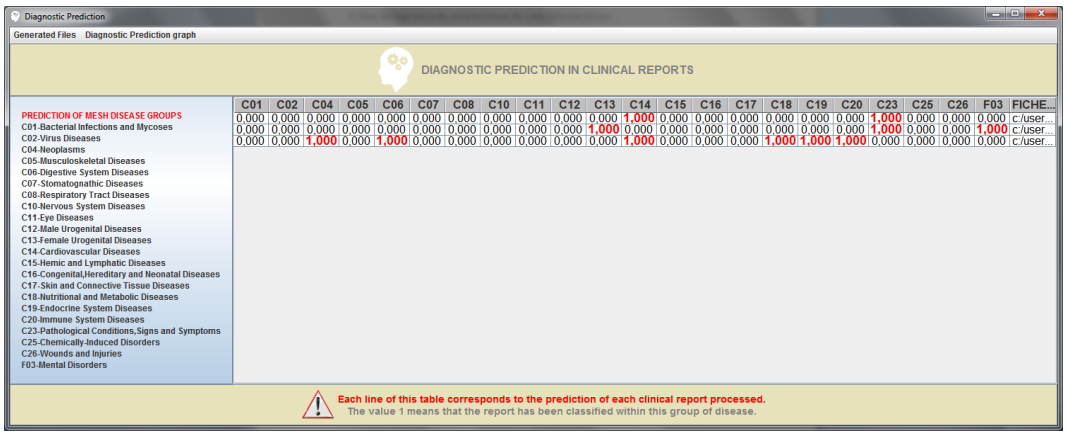


Figura 3.37. Clasificación diagnóstica masiva. Módulo +MiNerDoc

c) *Fase final de la predicción diagnóstica de múltiples informes. Ficheros test/training generados y representación gráfica de la clasificación.* Una vez obtenida la lista de las categorías diagnósticas MeSH en las que se ha clasificado cada uno de los informes clínicos seleccionados podemos ejecutar dos funcionalidades dentro del módulo +MiNerDoc. Una de ellas es la visualización de los ficheros test y training (extensión arff) que son utilizados para la construcción de la predicción diagnóstica masiva. A través de

Informe Clínico	Jerarquías diagnósticas en las que se ha clasificado
Discharge_report_1.txt	C14-Cardiovascular Diseases C23-Pathological Conditions, Signs and Symptoms
Discharge_report_2.txt	C13-Female Urogenital Diseases C23-Pathological Conditions, Signs and Symptoms F03-Mental Disorders
Discharge_report_3.txt	C04-Neoplasms C06- Digestive System Diseases C14- Cardiovascular Diseases C18- Nutricional and Metabolic Diseases C19- Endocrine System Diseases C20- Inmune System Diseases

Tabla 3.10. Ejemplo de Clasificación diagnóstica multietiqueta de múltiples informes clínicos. Opción +MiNerDoc.

esta opción será posible visualizar distintos ficheros generados durante el proceso de categorización, como el fichero inicial de bolsa de palabras. La segunda funcionalidad que permite MiNerDoc dentro del módulo de predicción masiva es la representación gráfica que nos permitirá visualizar el resultado global de la predicción diagnóstica de múltiples informes clínicos. Este gráfico nos dará una visión general del resultado de la clasificación múltiple, de esta forma será posible identificar visualmente que grupo o grupos de categorías diagnósticas son los más predominantes en el conjunto de informes clasificados, o que grupos de categorías diagnósticas concurren con otros grupos, o que porcentaje de informes se han clasificado en cada categoría diagnóstica predicha, etc. Un ejemplo de la representación gráfica obtenido de la clasificación de los tres informes clínicos puede visualizarse en la Figura 3.38. Como podemos observar en el ejemplo propuesto, de la lectura de este gráfico podemos extraer algunas conclusiones: *i)* los tres informes clínicos se han clasificado en 9 de las 22 categorías diagnósticas posibles,C04-Neoplasms,*C06-Digestive System Diseases,C13-Female Urogenital Diseases and Pregnancy Complications,C14-Cardiovascular Diseases,C18-Nutritional and Metabolic Diseases,C19-Endocrine System Diseases, C20-Immune System Diseases, C23-Pathological Conditions, Signs and Symptoms* y *F03-Mental Disorders;*

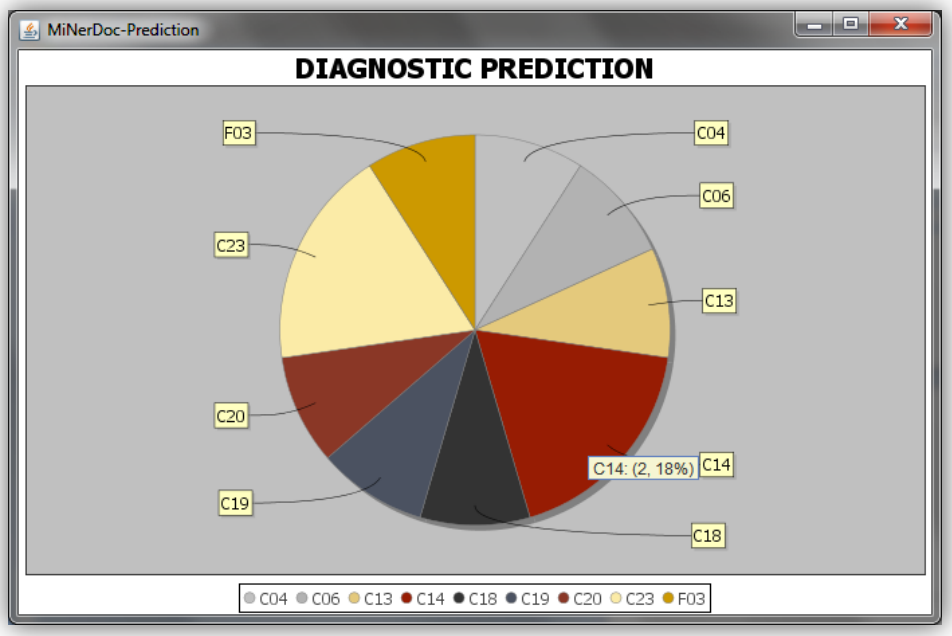


Figura 3.38. Gráfico generado despues de la predicción múltiple de informes clínicos (módulo +MiNerDoc).

ii) las categorías *C14-Cardiovascular Diseases* y *C23*-Pathological Conditions, Signs and Symptoms, son las más predominantes dentro de las clases en las que se han clasificado los informes clínicos; iii) si estuviéramos ante el procesamiento de múltiples informes clínicos de un único paciente podríamos analizar la concurrencia de patologías, en nuestro ejemplo confluirían patologías cardiovasculares junto con problemas de salud mental , neoplasias y problemas digestivos.

4

CASOS DE ESTUDIO

En este capítulo mostraremos el funcionamiento real de MiNerDoc presentando algunos casos de estudio donde analizaremos los aspectos más destacados del sistema de MT propuesto en esta tesis doctoral. Para mostrar estas funcionalidades se utilizaran una serie de informes de alta tomados de la colección MIMIC [55]. Esta base de datos ha sido seleccionada por tratarse de una colección de información clínica de acceso abierto para investigadores, anonimizada y creada en un entorno sanitario real (unidad de cuidados intensivos).

4.1. Caso I: Reconocimiento de entidades médicas y detección de factores de riesgo en un informe de alta.

En este primer caso de estudio, veremos cómo MiNerDoc detecta automáticamente los principales factores de riesgo asociados a una enfermedad (enfermedad cardíaca y respiratoria) en base a las entidades médicas identificadas en un informe de alta. Desde el editor de texto de MiNerDoc abriremos el informe clínico que deseemos analizar para obtener automáticamente las entidades médicas y los factores de riesgo (ver Figura 4.1).

INFORME_CLINICO .TXT

2958|||8951||||18099||||DISCHARGE_SUMMARY||||2009-01-10 00:00:00.0|||

Admission Date: [**2009-01-06**] Discharge Date: [**2009-01-10**]

Date of Birth: Sex:M

Service:

HISTORY OF PRESENT ILLNESS: The patient is a 71 year-old male with a history of coronary artery disease status post four vessel coronary artery bypass graft, congestive heart failure with an ejection fraction of 15%, diabetes, hypertension, chronic obstructive pulmonary disease, recent admission to [**Hospital 719**] in [**2008-07-30**] for congestive heart failure status post cardiac catheterization with a stent to saphenous vein graft to right coronary artery graft and status post stent to saphenous vein graft obtuse marginal graft in [**2004**]. Over the last two weeks the patient has had increasing shortness of breath, two pillow orthopnea, paroxysmal nocturnal dyspnea. The patient has also complained of dyspnea on exertion on walking 200 feet. The patient had a history of chronic renal insufficiency.

ALLERGIES: Shellfish and iodine.

HOSPITAL COURSE: The impression was that this was a 71 year-old male with a history of coronary artery disease, status post four vessel coronary artery bypass graft in [**2002**], recent stent to the saphenous vein graft to the right coronary artery in [**2008-07-30**] presenting with progressive shortness of breath. The patient was continued on aspirin, Lopressor 50 mg po b.i.d., Norvasc 5 mg po q.d. The patient had no chest pain during hospital stay.

The patient was initially placed on intravenous nitroglycerin drip.

The patient was continued on Lasix 80 mg po b.i.d., Monopril 20 mg po b.i.d., and was started on Aldactone 12.5 mg po q.d. The patient was also continued on Digoxin for congestive heart failure.

The patient had a history of atrial fibrillation and was on Amiodarone 200 mg po b.i.d. with a history of pacemaker placement for sick sinus syndrome. The patient had a V paced rhythm with baseline atrial fibrillation on his admission electrocardiogram. The Electrophysiology Service was consulted and given that the patient was in atrial fibrillation of an unknown duration, the patient was scheduled for DC cardioversion following a transesophageal echocardiogram, which was done during this hospital stay. The patient was successfully cardioverted and was subsequently normal sinus rhythm.

Dr. [**Last Name (STitle) 1473**].
[**First Name8 (NamePattern2) 491**] [**Last Name (NamePattern1) **], M.D. [**MD Number 1474**]
Dictated By:[**Dictator Info 1475**]
MEDQUIST36

D: [**2009-04-22**] 19:11
T: [**2009-04-23**] 08:54
JOB#: [**Job Number 1476**]
Signed electronically by: DR. [**First Name11 (Name Pattern1) 491**] [**Initials (NamePattern4) **] [**Last Name (NamePattern4) **] on: [**Doctor First Name 67**] [**2009-06-25**] 8:36 PM (End of Report)

Figura 4.1. Fragmento de Informe de alta de la colección MIMIC

Como hemos podido ver en el capítulo anterior, la tarea MER puede ejecutarse en MiNerDoc siguiendo dos enfoques: i) obtener la mayor cantidad de entidades médicas (independientemente del grado de concordancia con UMLS); ii) obtener las mejores entidades médicas (teniendo en cuenta el mayor grado de concordancia con UMLS). En este caso de estudio, nos centraremos en la segunda opción, obtención de aquellas entidades médica que tenga un alto grado de coincidencia con el término UMLS. Para obtener dichas entidades médicas, desde el menú principal pulsaremos la opción "*Best Mapping NER*" (ver Figura 4.2).

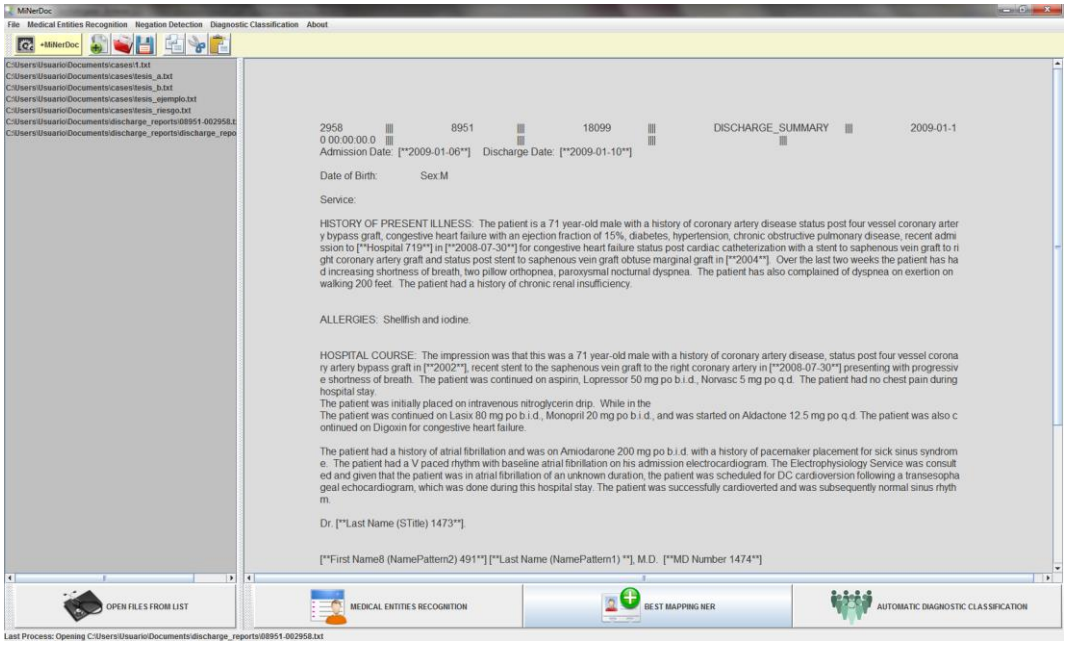


Figura 4.2. Opción Best Mapping NER de MiNerDoc

A continuación, aparecerá una ventana emergente donde los cinco grupos de entidades médicas, junto con las negaciones detectadas en el informe clínico, se presentaran al usuario (ver Figura 4.3). Como observamos en este ejemplo, algunas de las entidades médicas obtenidas tras el procesamiento del contenido textual de este informe de alta son las siguientes:

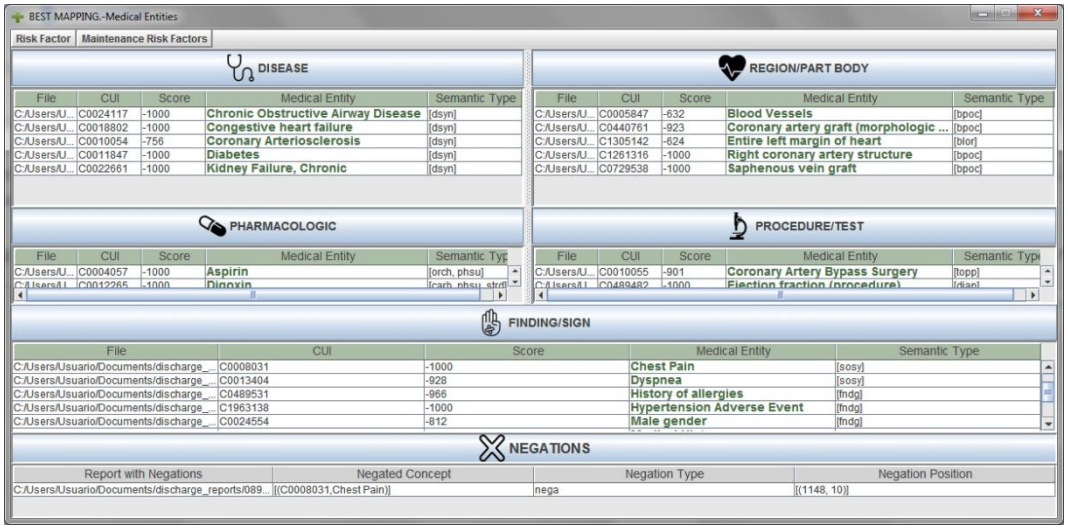


Figura 4.3. Reconocimiento de entidades médicas: opción Best Mapping NER

- Entidad "*Disease*": Congestive heart failure, Coronary Arteriosclerosis, Diabetes.
- Entidad "*Region/Part Body*": Coronary artery graft, Saphenous vein graft.
- Entidad "*Pharmacologic*": Aspirin, Digoxin, Lasix.
- Entidad "*Procedure/Test*": Cardiac Catheterization procedures, Coronary Artery Bypass surgery.
- Entidad "*Finding/sign*": Dyspnea, History of allergies, Hypertension.
- Negación encontrada: Chest pain.

Una vez identificadas las entidades médicas, MiNerDoc podrá obtener automáticamente los principales factores de riesgo de este informe clínico. Para ello, se requiere que el usuario defina previamente, en lenguaje natural, los riesgos potenciales o factores iniciales de riesgo asociados a un ámbito asistencial, enfermedades del corazón o enfermedades respiratorias. Este proceso sólo se realizará la primera vez que el usuario utilice la aplicación. El usuario registrará estos factores de riesgo iniciales a través de la pantalla donde se visualizan las entidades (ver Figura 4.3), pulsando sobre la opción de la barra de menú "*Maintenance Risk Factors*" → "*Add Initial Risk Factors*". Una vez

registrados estos factores iniciales (e.g. "*type 2 diabetes*" o "*low blood pressure*"), MiNerDoc expandirá estos términos considerando todas las variaciones terminológicas del mismo en base a la herramienta MetaMap y el metatesauro UMLS (ver Sección 3.4.4). A continuación, MiNerDoc podrá obtener los factores de riesgo de cualquier informe clínico gracias a la coincidencia entre los factores de riesgo expandidos y las entidades médicas obtenidas previamente del informe clínico (coincidencia del CUI de UMLS). Como hemos comentado, MiNerDoc permite detectar los factores de riesgo de dos ámbitos asistenciales, enfermedades del corazón y enfermedades respiratorias. En primer lugar realizaremos el proceso para obtener los factores de riesgo relacionados con las enfermedades del corazón, para ello, el usuario únicamente tendrá que pulsar sobre la opción del menú superior "*Risk Factor*" → "*Heart Disease Alerts*", y a continuación aparecerá una ventana emergente donde se marcarán en rojo los factores de riesgo detectados en el informe analizado (ver Figura 4.4). En este caso se han detectado cinco factores de riesgo, "*Chronic Obstructive Airway Disease*", "*Coronary Arteriosclerosis*", "*Diabetes*", "*Hypertension adverse event*" y "*Dyspnea*". Si optamos por obtener los factores de riesgo relacionados con las enfermedades respiratorias, sólo tendríamos que seleccionar la opción del menú superior "*Risk Factor*" → "*Respiratory Disease Alerts*", y aparecerá una ventaja emergente con los factores de riesgo detectados (ver Figura 4.5), como por ejemplo, *dyspnea o diabetes*. Previamente el usuario ha tenido que definir (en lenguaje natural y sólo una única vez) los factores de riesgo potenciales para este ámbito asistencial.

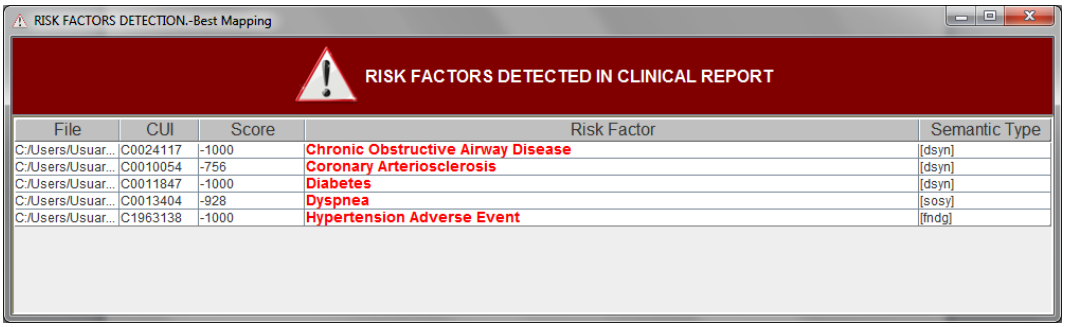


Figura 4.4. Factores de riesgo encontrados en un informe clínico (ámbito enfermedades del corazón)

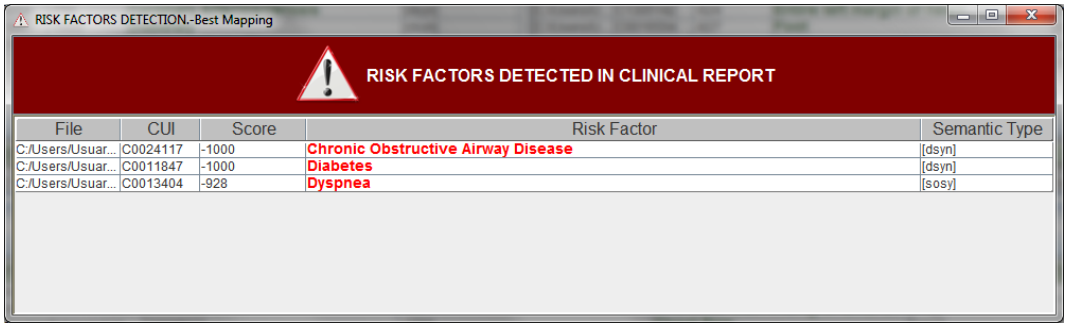


Figura 4.5. Factores de riesgo encontrados en un informe clínico (ámbito enfermedades respiratorias)

4.2. Caso II: Clasificación diagnóstica automática de un informe de alta.

En el segundo caso de estudio mostraremos como MiNerDoc, de forma sencilla y rápida, lleva a cabo la tarea de asignar automáticamente códigos de diagnósticos normalizados (22 descriptores MeSH asociados con enfermedades) a un informe de alta. MiNerDoc permite realizar esta tarea de dos formas distintas: i) seleccionando paso a paso todas las fases de MT para llegar a la clasificación final; ii) o realizando el proceso completo de clasificación con un simple click. En este caso de estudio, se seleccionó la segunda opción, para ello comenzaremos cargando desde el menú principal de MiNerDoc el informe de alta utilizado en el caso anterior (ver Figura 4.1) para poder obtener la predicción diagnóstica. Una vez abierto el informe pulsaremos sobre el botón inferior del menú principal *"Automatic Diagnostic Classification*" (ver Figura 4.6) y una ventana emergente aparecerá en la pantalla donde tendremos que pulsar el botón inferior *"Execute Complete Prediction Process*" (Figura 4.7). A continuación, aparecerá una ventana (Figura 4.8) en la que la predicción de los grupos de diagnósticos MeSH se mostraran marcados en rojo. Como observamos, MiNerDoc clasificó el informe clínico de este caso de estudio en cinco grupos de diagnóstico MeSH*, C08-Respiratory Tract*

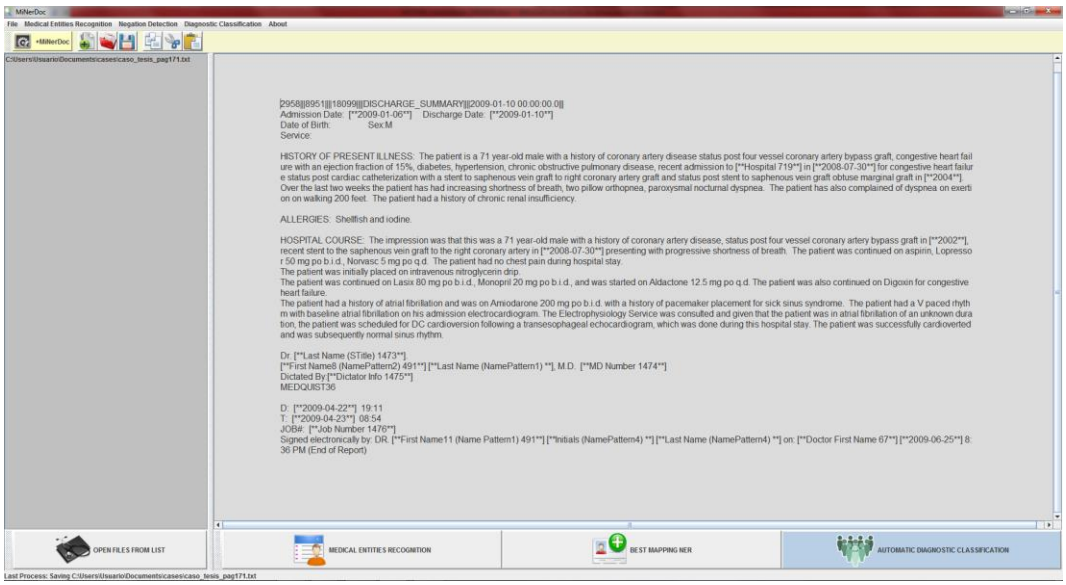


Figura 4.6. Opción "Automatic Diagnostic Classification" desde menú principal de MiNerDoc

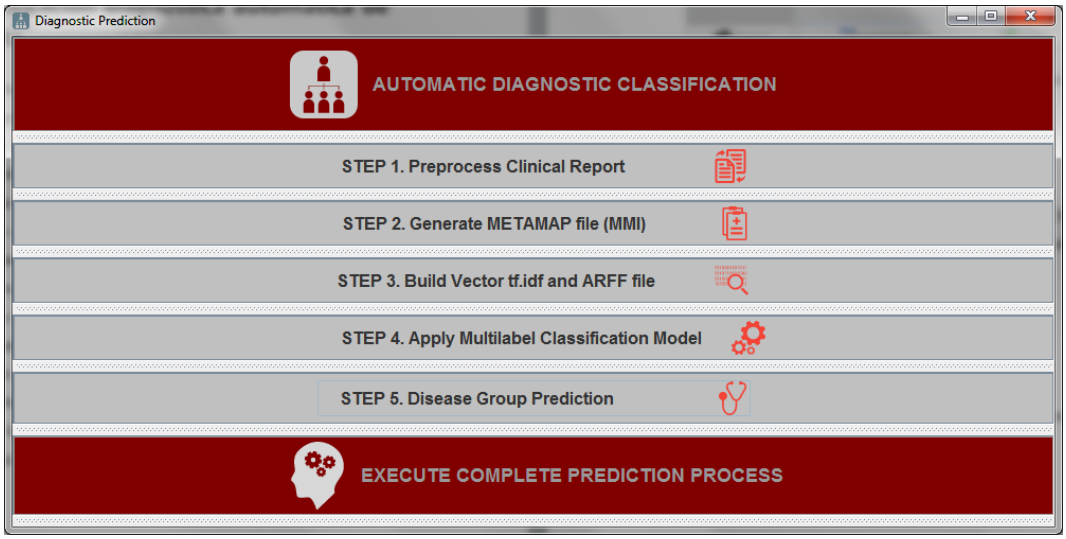


Figura 4.7. Ventana que inicia el proceso de clasificación diagnóstica de un informe clínico.

Diseases, C14-Cardiovascular Diseases, C18-Nutritional and Metabolic Diseases y *C19-Endocrine System Diseases*. Por tanto, la predicción de los grupos de diagnósticos MeSH de este informe clínico se ha centrado en las enfermedades del aparato respiratorio, las

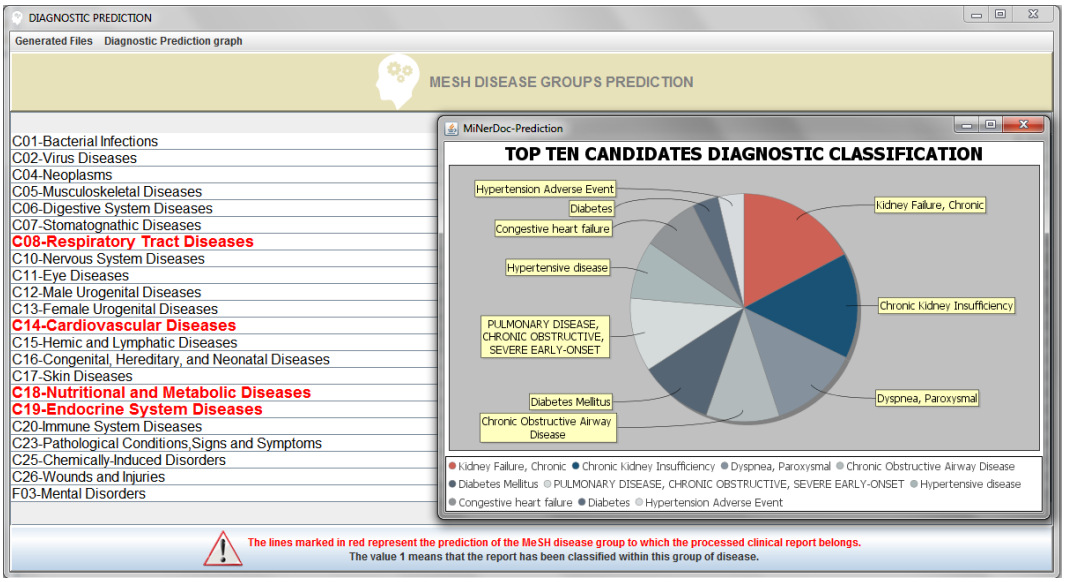


Figura 4.8. Clasificación diagnóstica automática de un informe clínico. Gráfico del top-ten de términos candidos.

enfermedades cardiovasculares, las enfermedades nutricionales y metabólicas y por último, en la enfermedades del sistema endocrino. Con el objetivo de aportar valor a los resultados obtenidos, MiNerDoc permite representar gráficamente los diez términos candidatos principales (conceptos relacionados con diagnósticos o síntomas más destacados) que han formado parte del proceso de clasificación y que han obtenido un mayor nivel de puntuación según la función MMI de MetaMap [240] (a mayor puntuación mayor relevancia del concepto). En nuestro ejemplo, algunos de los términos top-ten provenientes del informe de alta analizado con una mayor puntuación de ranking según la salida MMI de MetaMap y por tanto, con un mayor nivel de concordancia con UMLS son, entre otros: "*Kidney Failure, Chronic*", "*Chronic Kidney Insufficiency*", "*Dyspnea, Paroxysmal*", "*Chronic Obstructive Airway Disease*", "*Diabetes Mellitus*", etc.

4.3. Caso III: Clasificación diagnóstica automática de una colección de informes de alta.

En el tercer caso de estudio analizaremos una de las funcionalidades más destacadas y útiles de MiNerDoc, la clasificación diagnóstica de un conjunto de informes de alta con tan sólo pulsar un click. Para ello partiremos de un conjunto de 10 fragmentos de informes de altas tomados de la colección MIMIC[55]. Para llevar a cabo esta tarea seguiremos los siguientes pasos: i) situaremos los 10 informes de alta (o el número de informes de los que queramos realizar la predicción diagnóstica) en un directorio que MiNerDoc marca por defecto; ii) a continuación, pulsaremos sobre la opción del menú principal *+MiNerDoc*, aparecerá una ventana emergente y sólo será necesario pulsar sobre el botón inferior ("*Execute complete process for diagnostic predicction*") y el proceso de clasificación masivo se llevará a cabo (ver Figura 4.9). Algunos de los fragmentos de informes de alta seleccionados de la colección MIMIC [55] para llevar a cabo el proceso de predicción diagnóstica masiva pueden verse en la Figura 4.10.

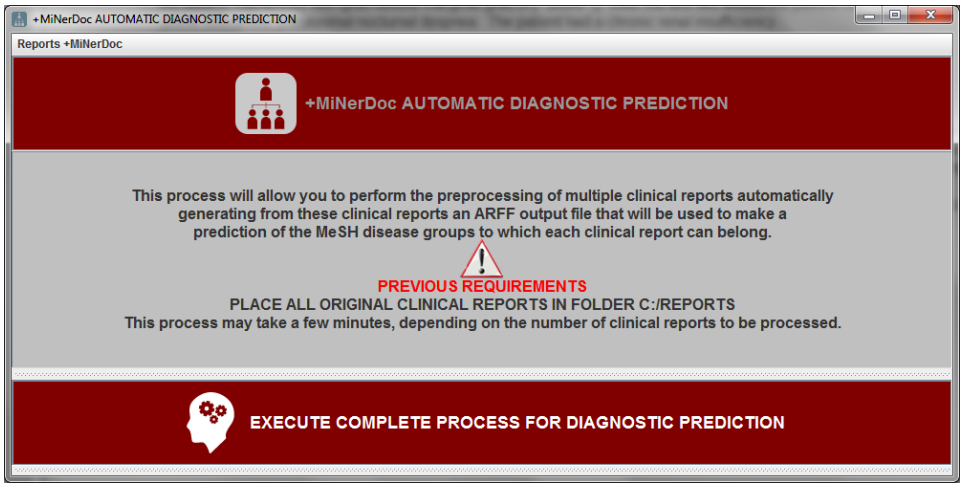


Figura 4.9. Clasificación diagnóstica automática masiva

INFORME CLINICO 1.TXT

Admission Date: [**2011-10-06**] Discharge Date: [**2011-10-17**]
Date of Birth: [**1935-03-29**] Sex: M
Service: Medicine

CHIEF COMPLAINT: Admitted from rehabilitation for hypotension (systolic blood pressure to the 70s) and decreased urine output.
HISTORY OF PRESENT ILLNESS: The patient is a 76-year-old male who had been hospitalized at the [**Hospital 3723**] from [**09-27**] through [**10-05**] of 2002 after undergoing a left femoral-AT bypass graft and was subsequently discharged to a rehabilitation facility.
On [**2011-10-06**], he presented again to the [**Hospital 3109**] after being found to have a systolic blood pressure in the 70s and no urine output for 17 hours.
A Foley catheter placed at the rehabilitation facility yielded 100 cc of murky/brown urine. There may also have been purulent discharge at the penile meatus at this time.
On presentation to the Emergency Department, the patient was without subjective complaints. In the Emergency Department, he was found to have systolic blood pressure of 85. He was given 6 liters of intravenous fluids and transiently started on dopamine for a systolic blood pressure in the 80.s

INFORME CLINICO 2.TXT

HOSPITAL COURSE: The patient was admitted to [**Hospital1 **] [**First Name (Titles) 2065**] [**Last Name (Titles) 2066**] Psychiatric Services for further evaluation, as well as safety and stabilization. The patient completed a short course of Levofloxacin begun in the Emergency Department for presumed urinary tract infection.
After consultation with the team, the patient agreed to begin Citalopram for his depression, as well as reported anxiety.
She was also begun on a very low dose of Clonazepam as needed for his complaints of generalized anxiety. She had stated that this had benefitted her in the past when in stressful situations. She tolerated all of the medications well. She reported eating and sleeping better following admission.
Throughout her stay, she denied suicidal thoughts and exhibited no unsafe behaviors. She did appear anxious about leaving however. A full explanation for her motivations for admission was not fully elucidated while talking to the patient. There appeared to be some characterologic component to her presentation; however, the patient refused to complete a [**State 2067**] Multiphasic Personality Inventory, impairing our diagnostic capability.

INFORME CLINICO 3.TXT

HISTORY OF PRESENT ILLNESS: The patient is a 79 year-old man with a past medical history significant for coronary artery disease status post coronary artery bypass graft in [**2005**] as well as at and insulin dependent diabetes mellitus found to have hepatic flexure tumor positive for dysplasia. The patient presented with blood in his stool and on colonoscopy the lesion was noted in the hepatic flexure of the colon.

HOSPITAL COURSE: The patient was admitted on [**2011-02-07**] and taken to the Operating Room for a right colectomy. The patient tolerated the procedure well and was transferred to the PACU and then to the floor in stable condition. His postoperative course was essentially uneventful. He was seen by cardiology as well as [**Last Name (un) 411**] for monitoring of his cardiac medications as well as his insulin. Due to low blood pressure postoperative the patient's Lisinopril was discontinued and his Carvedilol dose was halved to 12.5 po b.i.d. The patient's diet was advanced and on postop day number six the patient was ready for discharge when he fell getting out of the bathroom. He did not hit his head. He had no loss of consciousness. Vital signs were stable. Given this event the patient was seen by physical therapy for clearance and on postoperative day seven [**2011-02-14**] the patient was discharged home in stable condition.

Figura 4.10. Ejemplo de informes de la colección MIMIC clasificados según opción +MiNerDoc

En la Figura 4.11 se recoge el resultado final de la clasificación conjunta de los 10 informes de alta, en cada fila se recogerá el resultado de la predicción diagnóstica de cada informe de alta y en cada columna se recoge el grupo de diagnóstico MeSH en el que se ha clasificado dicho informe, así el primer informe se ha clasificado en los grupos diagnósticos C14 (enfermedades cardiovasculares) y C23 (condiciones patológicas, signos y síntomas), el segundo informe en los grupos diagnósticos C13 (enfermedades urogenitales femeninas y complicaciones del embarazo), C23 (condiciones patológicas, signos y síntomas) y F03 (enfermedad mental).

MiNerDoc permitirá representar gráficamente el resultado de esta predicción final (ver Figura 4.12), como podemos observar la categoría diagnóstica predominante es la C14-Cardiovascular Diseases, seguida del C23-Pathological Conditions, Sign and Symptoms (es decir, las enfermedades cardiovasculares y el grupo que recoge las condiciones, signos y síntomas patológicos). Los informes se han clasificado con un mínimo de una categoría diagnóstica hasta un máximo de 6 categorías diagnósticas.

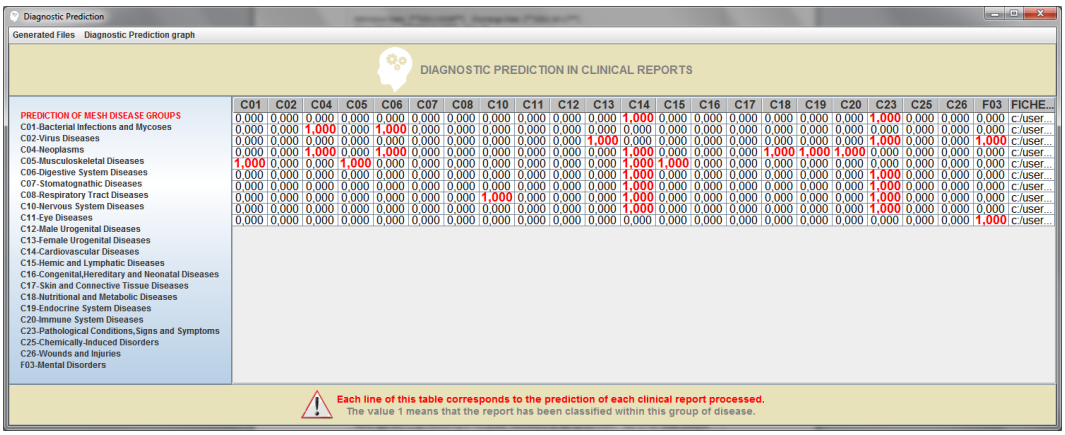


Figura 4.11. Resultado final de la clasificación masiva de 10 informes de la colección MIMIC clasificados según opción +MiNerDoc

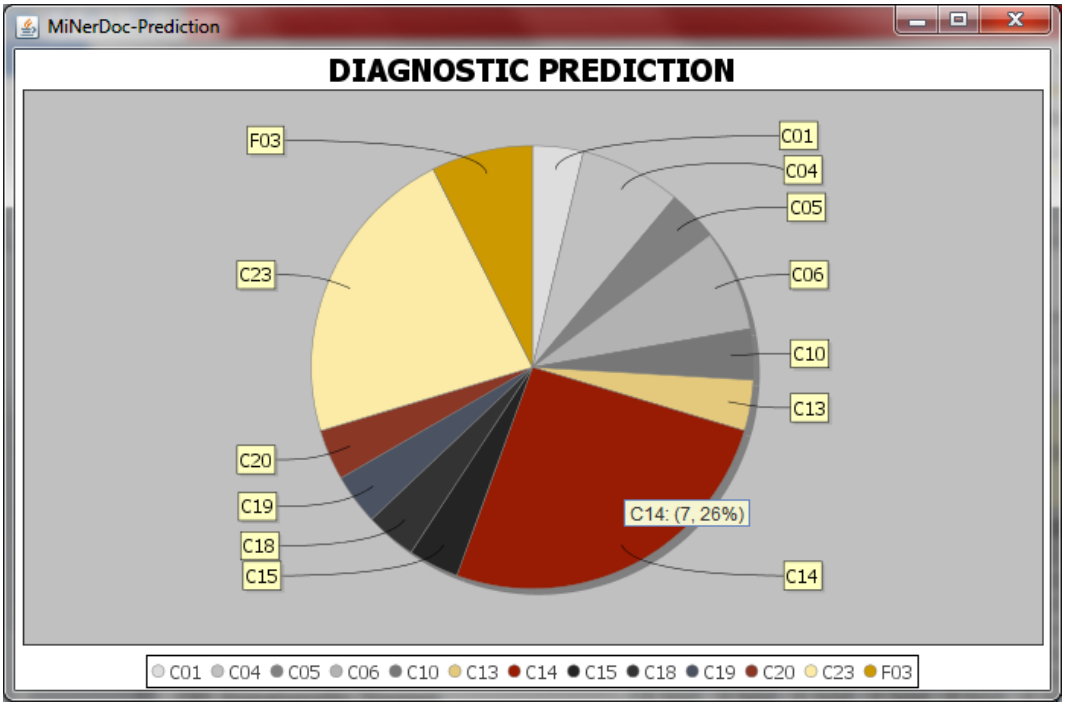


Figura 4.12. Representación gráfica del resultado final de la clasificación de 10 informes de la colección MIMIC clasificados según opción +MiNerDoc

Con la funcionalidad de la predicción masiva, MiNerDoc facilita la ardua y compleja tarea de la asignación de códigos de diagnóstico a una colección de informes de alta, realizando el proceso de forma ágil, intuitiva y con un ahorro importante en tiempo y recursos. Además, esta funcionalidad puede aportar un conocimiento adicional que apoye al clínico en la tarea de toma de decisiones, ya que a través de esta utilidad será posible analizar múltiples informes clínicos (de distintas especialidades, distintos periodos de tiempo, etc) de un único paciente y así se podrá determinar, con un simple click y en un simple vistazo, que patologías son las más frecuentes, sobre cuales es necesario actuar o si pueden existir patologías concomitantes.

5

Experimentación

El objetivo de este capítulo es describir el análisis experimental diseñado para evaluar los dos subsistemas que constituyen la aplicación MiNerDoc, reconocimiento de entidades médicas y clasificación diagnóstica. Cada uno de estos subsistemas se evaluará a través de una amplia gama de experimentos que se detallarán en este capítulo. En primer lugar evaluaremos el sistema de reconocimiento de entidades médicas (MER) desarrollado en esta tesis doctoral. Para ello, hemos realizado una comparativa del desempeño de nuestro sistema MER con dos conocidos sistemas open-source para el reconocimiento de entidades nombradas en textos clínicos. En segundo lugar, evaluaremos la metodología propuesta para mejorar el rendimiento de la clasificación diagnóstica automática (CDA), a la que hemos denominado *diagnostic Classification with Semantic Enrichment* (dCSE), y determinaremos cuál es la mejor combinación de técnicas MT (*stemming*, *stop-words*, distintos niveles de granularidad en la tokenización) que hacen mejoran el desempeño de dicha tarea. Por último, evaluaremos el desempeño del método de aprendizaje multietiqueta (MLL, por sus siglas en inglés) [229, 231] empleado en el sistema CDA de MiNerDoc. Para ello, serán

comparados once métodos MLL del estado del arte para determinar cuál de ellos ofrece un rendimiento predictivo óptimo en la tarea de clasificación propuesta. Uno de los objetivos principales se centrará en demostrar, en base a diferentes análisis estadísticos, que la utilización de herramientas externas específicas del dominio médico (como MetaMap y UMLS) pueden aportar un valor añadido que hacen mejorar el proceso final de extracción de entidades médicas y el proceso de asignación automática de códigos de diagnóstico normalizados. A continuación, resumiremos el planteamiento de cada uno de los experimentos propuestos, las preguntas de investigación que pretendemos responder y el objetivo principal de cada uno de ellos.

Experimento 1. Evaluar el desempeño del sistema MER incluido en MiNerDoc

Preguntas de investigación

- ¿Qué desempeño puede tener un sistema MER basado en diccionarios?
- ¿Es competitivo nuestro sistema MER frente a otros sistemas basados en AA o en sistemas híbridos?
- ¿Qué errores relacionados con la detección de entidades médicas se detectan en el sistema MER propuesto?
- ¿Cuáles son los rasgos diferenciadores del sistema MER de MiNerDoc con respecto a los dos sistemas evaluados?

Objetivo principal

Comparar el desempeño de nuestro sistema MER (incluido en la aplicación MiNerDoc) con dos conocidos sistemas para el reconocimiento de entidades médicas, denominados CliNER y CLAMP.

Experimento 2. Evaluar qué metodología (dCSE versus Baseline) y qué parametrización (aplicación de distintas técnicas de MT) hacen mejorar el desempeño del sistema CDA de MiNerDoc.

Preguntas de investigación

- ¿Puede la MT resolver con calidad y eficacia la categorización automática de información textual dentro del ámbito de la codificación diagnóstica multietiqueta?
- ¿La utilización de herramientas externas de conocimiento, como MetaMap y el metatesaurus UMLS, pueden mejorar la calidad de la categorización automática de informes clínicos?
- ¿Qué papel juega la selección de características o rasgos de cada colección en la eficacia de la clasificación diagnóstica multietiqueta?
- ¿Cuál es la parametrización de técnicas de MT (*stemming*, distintas granularidades en la tokenización, enriquecimiento semántico, eliminación stop-words) que aporta una mejora en la categorización diagnóstica?

Objetivos principales

Determinar si el enriquecimiento semántico proporcionado por la herramienta MetaMap y el metatesaurus UMLS aumenta el desempeño en la tarea de codificación diagnóstica multietiqueta**.** Evaluar diferentes técnicas de MT para formar los distintos vocabularios (parametrizaciones) y establecer que combinación hace mejorar la tarea de clasificación diagnóstica.

Experimento 3. Determinar qué método de aprendizaje multietiqueta ofrece un mejor resultado en la tarea CDA

Preguntas de investigación

- ¿Qué métodos MLL mejoran la eficacia de la categorización diagnóstica?
- ¿Qué método de clasificación multietiqueta es óptimo para aumentar el desempeño de la categorización diagnóstica en MiNerDoc?

Objetivo principal

Evaluar el rendimiento de 11 métodos de clasificación multietiqueta del estado del arte y determinar cuál de ellos ofrece un rendimiento óptimo en la tarea de clasificación diagnóstica de informes de alta.

En las siguientes secciones analizaremos detalladamente todos los procesos llevados a cabo en cada uno de los experimentos realizados, detallaremos la configuración aplicada en cada uno de los tres experimentos, recogeremos los resultados obtenidos y por último realizaremos una discusión de los mismos.

5.1. Experimentación 1: evaluar el desempeño del sistema MER

El principal objetivo de este primer experimento es determinar el rendimiento del sistema MER incluido en el sistema MiNerDoc, comparando dicho rendimiento con dos conocidos sistemas para la extracción de entidades médicas denominados CliNER[46] y CLAMP[47]. CliNER está basado en un enfoque AA y CLAMP sigue un enfoque híbrido (AA y reglas). A continuación, describiremos en primer lugar la configuración necesaria para abordar este primer experimento (creación del corpus semántico y definición de métricas empleadas) y en segundo lugar se recogerán los resultados de la experimentación llevada a cabo.

5.1.1. Configuración experimental

Para evaluar el rendimiento del sistema MER desarrollado en esta tesis doctoral, se ha llevado a cabo la construcción de un corpus semántico anotado manualmente por un médico experto en documentación clínica del Hospital Universitario Reina Sofía. Este corpus constituirá el Gold-Standard que será utilizado para medir el rendimiento de nuestro sistema MER frente a otros sistemas ampliamente conocidos en el ámbito del reconocimiento de entidades médicas. El corpus anotado manualmente está formado

[46] http://text-machine.cs.uml.edu/cliner/
[47] https://clamp.uth.edu/index.php

```
HISTORY OF PRESENT ILLNESS:  The patient is an 88 year
old female with coronary artery disease, congestive heart
failure and diabetes mellitus who presented with fever,
abdominal pain after being found down at her nursing
home.  Her history patient is a resident at [**Hospital
192**] who was found status post questionable fall the
morning of admission and was noted to have a left-sided
weakness without head trauma or loss of consciousness.
```

Figura 5.1.- Ejemplo de fragmento original que formará parte del corpus anotado

por una colección de 20 fragmentos de informes de altas tomados de la colección MIMIC[55], un ejemplo de estos textos clínicos puede visualizarse en la Figura 5.1.

Los textos clínicos han sido anotados teniendo en cuenta tres tipos de entidades diferentes, enfermedades o problemas, test o procedimientos y medicación o tratamientos. La entidad enfermedad se marcó entre las etiqueta *<problema></problema>*, la entidad test o procedimiento diagnóstico se marcó con la etiqueta *<procedimiento></procedimiento>* y por último, la entidad médica medicación o tratamientos se etiquetó como *<tratamiento></ tratamiento>.*

Se evaluará, en primer lugar, para cada fragmento de informe de alta y cada sistema MER evaluado, el número total de entidades médicas reconocidas y el número de entidades correctamente etiquetadas en relación al Gold-Standard. Un ejemplo de un fragmento de informe de alta etiquetado manualmente se recoge en la Figura 5.2.

```
Routine laboratories were sent in the Emergency Department
and the patient was found to have a <procedimiento> chest x-
ray  </procedimiento> consistent  with a  <problema>  right
middle lobe pneumonia </problema>. He was also febrile in
the  Emergency  Room.  His  mental  status  improved
with <tratamiento> Tylenol </tratamiento>.
```

Figura 5.2.- Ejemplo de informe etiquetado manualmente por anotador (Gold-Standard).

Es importante resaltar que el corpus es limitado debido a la dificultad en encontrar expertos anotadores que realicen la ardua y compleja labor de la anotación manual.

Las métricas de evaluación utilizadas para llevar a cabo esta experimentación son las métricas denominadas "*exact match*" [56] o de "coincidencia exacta", ampliamente utilizadas en la evaluación de sistemas de reconocimiento de entidades. Estas medidas fueron presentadas en las conferencias CoNLL (*Conference on Computational Natural Language Learning*) donde consideraron que una entidad nombrada es correcta únicamente cuando existe una coincidencia exacta entre la entidad y el término etiquetado en el Gold-Standard. Siguiendo esta norma, hemos realizado una evaluación global del sistema MER propuesto, analizando cada informe de alta (20 seleccionados) frente a cada sistema MER evaluado (MiNerDoc, CliNER y CLAMP) y considerando que una entidad médica se puntuará como correcta solo si se da una coincidencia exacta entre dicha entidad y la entidad del Gold-Standard. Las métricas exact-match empleadas han sido Precisión, Recall y FMeasure (ver Tabla 5.1). La precisión hace referencia al porcentaje de entidades nombradas encontradas por el sistema que son correctas. La métrica recall hace referencia al porcentaje de entidades nombradas presentes en el corpus que son encontradas por el sistema. Por último, la medida FMeasure es la media armónica entre las métricas precisión y recall.

Métricas de evaluación sistema MER

$$Precisión = \frac{nº\ de\ etiquetas\ correctas\ dadas\ por\ el\ sistema}{nº\ de\ etiquetas\ identificadas\ por\ el\ sistema}$$

$$Recall = \frac{nº\ de\ etiquetas\ correctas\ dadas\ por\ el\ sistema}{nº\ de\ etiquetas\ del\ Gold-Standard}$$

$$FMeasure = \frac{2 * Presición * Recall}{Precisión + Recall}$$

Tabla 5.1. Métricas para evaluación de sistemas MER.

Para profundizar en el rendimiento de cada sistema MER, se analizarán globalmente algunos de los principales errores que se detectan habitualmente en la mayoría de los sistemas de reconocimiento de entidades y que son de especial relevancia en el ámbito médico. Los errores que se analizaran a nivel global en el procesamiento de cada informe de alta en los tres sistemas MER evaluados son:

- *Errores en la delimitación de la entidad médica*: uno de los errores más comunes en los sistemas de reconocimiento de entidades es la inexactitud en delimitación de la entidad. Algunos sistemas no detectan correctamente el nombre de la entidad e incorporan algunos tokens más a la entidad (artículos, pronombres, adjetivos, etc).
- *Errores en la detección de acrónimos:* la detección de acrónimos es una de las tareas de mayor importancia en el ámbito de la Medicina, debido a que su uso en los textos clínicos es muy extendido por la mayoría de los profesionales sanitarios. Las dificultades para su detección son elevadas, los acrónimos son muy dependientes del dominio y su uso aumenta los problemas de sinonimia y polisemia.
- *Errores en la resolución de acrónimos:* además de detectar un acrónimo, un buen sistema MER deber ser capaz de resolver e interpretar las siglas encontradas (por ejemplo el acrónimo *CHF* debe resolverse como *Congestive Heart Failure*). Para ello, el sistema debe enfrentarse a los problemas de desambiguación para encontrar el término más adecuado según el contexto en el que se encuentre la entidad.
- *Errores en la detección de negaciones:* como hemos comentado a lo largo de esta tesis doctoral, es de gran importancia en el ámbito clínico la detección de los conceptos negados, ya que la interpretación de un texto puede cambiar ante la detección de una afirmación o una negación de la entidad médica.

- *Errores en la clasificación del tipo de entidad:* la incorrecta asignación de la categoría a la que pertenece una entidad médica también es una fuente de errores en la mayoría de los sistemas de reconocimiento de entidades. Algunos sistemas MER pueden identificar erróneamente algunos medicamentos como procedimientos o viceversa, o pueden clasificar la misma entidad en dos categorías diferentes.
- *Detección incorrecta de entidades médicas (falsos positivos y falsos negativos):* otro de los problemas que pueden darse en los sistemas MER se centra en la falsa detección de entidades médicas que realmente no existían en el texto clínico del que se partía o por el contrario, algunos sistemas MER pueden dejar de detectar entidades médicas que realmente si existían en el informe clínico.

5.1.2. Resultados

A continuación, se mostrará la tabla de resultados obtenida del análisis de los tres sistemas MER evaluados frente al Gold-Standard (ver Tabla 5.2) y los errores detectados en el reconocimiento de las entidades nombradas (ver Tabla 5.3). Como podemos observar en base a los resultados obtenidos en esta experimentación, considerando el total de 186 entidades anotadas en el Gold-Standard, los tres sistemas MER evaluados han identificado un número superior de entidades totales, CLAMP reconoció 218 entidades médicas, CliNER 206 entidades y MiNerDoc un total de 204 entidades. Con respecto al número de entidades médicas correctamente reconocidas, el sistema CLAMP encontró 161 entidades, CliNER detectó 158 entidades y MiNerDoc identificó 159 entidades médicas. La principal diferencia radica en el número de entidades médicas que cada sistema MER identifica incorrectamente, así el sistema que identifica un número menor de entidades incorrectas es MiNerDoc (45 entidades), seguido del sistema CliNER (48 entidades incorrectas) y por último, el que obtiene más entidades

incorrectas es el sistema CLAMP (57 entidades). Basándonos en las métricas *exact-match*, el sistema MER incluido en MiNerDoc obtiene el primer puesto en las métricas precisión (77.94%) y FMeasure (81.54), y el segundo puesto en cuanto a la métrica Recall (85.48%).

Sistemas MER evaluados	Total de entidades reconocidas	Entidades correctas	Entidades incorrectas	Precision	Recall	FMeasure
CLAMP	218	161	57	73.85	**86.56**	79.70
CliNER	206	158	48	76.70	84.95	80.61
MiNerDoc	204	159	45	**77.94**	85.48	**81.54**
Gold-Standard: 186 entidades médicas anotadas sobre 20 informes de altas analizados de la colección MIMIC						

Tabla 5.2. Resultados entidades médicas reconocidas por los sistemas MER evaluados. Los mejores resultados se destacan en negrita.

En cuanto a los resultados globales sobre los principales errores que suelen presentarse en los sistemas MER, observamos en la Tabla 5.3 los resultados obtenidos de los tres sistemas evaluados:

Errores sistemas MER	CLAMP	CliNER	MiNerDoc
Errores delimitación entidad	**Si**	**Si**	No
Errores detección Acrónimos	No	No	No
Errores resolución Acrónimos	**Sí**	**Sí**	No
Errores detección Negaciones	No	**Sí**	No
Errores en la clasificación	**Sí**	No	No
Falsos negativos y Falsos positivos	**Sí**	**Sí**	**Sí**

Tabla 5.3. Errores encontrados en los sistemas MER evaluados

El detalle de los errores más representativos que se han presentado en los distintos sistemas MER evaluados (CLAMP, CliNER, MiNerDoc) han sido los siguientes:

- **Errores en la delimitación de la entidad.** Uno de los principales problemas detectados en el sistema CLAMP es que no hace una extracción totalmente correcta de la entidad, hay algunos errores en la delimitación de la entidad, incorporando determinantes a la verdadera entidad. Así, por ejemplo, el sistema CLAMP extrae la entidad *"His antihypertensive medications*" en lugar de *"antihypertensive medications*". En algunos casos, se produce una duplicidad en la extracción de la entidad debido principalmente al problema en la delimitación de la entidad, extrayendo por ejemplo dos entidades como *"stent"* y *"a stent",* en lugar de una. En cuanto al sistema CliNER, existe un mayor número de errores detectados en cuanto a la delimitación de la entidad que el resto de los sistemas evaluados. Así, por ejemplo, igual que ocurría con el sistema CLAMP, añade a la entidad detectada pronombres o artículos innecesarios, como "her Coumadin" en lugar de "Coumadin". En el sistema MER de MiNerDoc no se genera este error, debido a que la selección de la entidad médica se realiza a través de los tipos semánticos de UMLS proporcionados gracias a la herramienta MetaMap, esto se traduce en un mejor ajuste en la calidad de la selección de términos ya que sólo son seleccionados aquellos candidatos en los que el grado de similitud entre la palabra de la colección y cada uno de los candidatos obtenidos del metatesauro UMLS es totalmente exacto.
- **Errores en la detección de acrónimos.** Los sistemas CLAMP, CliNER y MiNerDoc realizan correctamente la detección de los acrónimos existentes en los informes de alta analizados.
- **Errores en la resolución de acrónimos.** El sistema CLAMP realiza correctamente la detección de acrónimos aunque no realiza la conversión del acrónimo por la descripción completa de dicho término, hecho que puede quitar valor a la

detección de la entidad si el usuario final no parte del conocimiento previo de ese acrónimo. Al igual que CLAMP, el sistema CliNER también realiza correctamente la detección de acrónimos aunque tampoco realiza la interpretación de dicho término. En cuanto al sistema MER propuesto, la detección ha sido correcta (al igual que el resto de sistemas MER evaluados), pero incorpora una importante ventaja con respecto al resto, la conversión automática del acrónimo por su significado completo, por ejemplo, en lugar de extraer la entidad *"CHF"* como tal, extrae la interpretación completa de dicho acrónimo (*"Congestive heart failure").*

- **Errores en la detección de negaciones.** El sistema CLAMP detecta correctamente las negaciones de las entidades médicas, marcando la entidad negada como *"absent". E*l sistema CliNER no detecta las entidades negadas, hecho que puede dar lugar a una lectura errónea del informe clínico, ya que no es lo mismo afirmar que negar la existencia de una enfermedad o un tratamiento. El sistema MER de MiNerDoc detecta correctamente las negaciones de las entidades médicas encontradas en cada texto clínico.
- **Errores en la clasificación del tipo de entidad.** El sistema CLAMP ha presentado algunos problemas en cuanto a la asignación de la categoría a la que pertenece la entidad médica, produciéndose asignaciones del mismo término en dos categorías diferentes. Así, por ejemplo, los términos *"stent stenosis",* se han clasificado por el sistema CLAMP dentro de la categoría tratamiento y dentro de la categoría problema al mismo tiempo. Sin embargo, no se han detectado problemas en cuanto a la asignación de la clase a la que pertenece cada entidad por el sistema CliNER. En cuanto al sistema MER de MiNERDOC la asignación de la categoría a cada entidad médica es correcta, tratándose de una clasificación fiable ya que se basa en el metatesauro UMLS.
- **Detección de entidades incorrectas (falsos positivos y falsos negativos).** El número de entidades médicas no detectadas por CLAMP y que sí se encuentran

encuentran en el Gold-Standard, es decir los falsos negativos detectados, es muy bajo, por este motivo el sistema CLAMP refleja el valor de Recall más alto. En cuanto a los falsos positivos, es decir, aquellas entidades que encuentra el sistema CLAMP pero que no se encuentran en el Gold-Standard, podemos observar que existe un alto número de casos, por ejemplo, el sistema ha detectado el siguiente fragmento *"a positive stress test in" co*mo un problema o enfermedad cuando en realidad según el Gold-Standard la entidad detectada debería ser "*stress test*", dentro de la categoría de procedimiento. Otros falsos positivos encontrados son, por ejemplo, la detección de palabras genéricas como "*evaluation*" que son marcadas por CLAMP como una entidad del tipo procedimiento. El número de entidades no encontradas por CliNER y que sí se encuentran en nuestro Gold-Standard (falsos negativos) es algo más alto que en el resto de sistemas evaluados, por este motivo este sistema refleja el valor de Recall más bajo. En cuanto a los falsos positivos, podemos observar que existe un número alto de casos, por ejemplo, el sistema ha marcado el siguiente fragmento *"<problem> diabetes </problem> type 2",* cuando en realidad según el Gold-Standard la entidad detectada debería ser "*diabetes type 2*". Otros falsos positivos encontrados en el sistema CliNER, son la detección de palabras genéricas como *"<treatment>volume</treatment>*" que son marcadas por CLiNER como entidades cuando realmente no lo son. En cuanto al sistema MER de MiNerDoc, existen algunas entidades que no han sido encontradas por el sistema MER propuesto y que sí se encuentran en el Gold-Standard (falsos negativos), este hecho se refleja en el valor de Recall (segunda posición). También se han detectado varios falsos positivos, así, por ejemplo, el sistema detecta variaciones semánticas de algunos conceptos, como *"Back Pain"* y *"Back Pain Adverse Event"* cuando en el Gold-Standard sólo está anotada la entidad *"back pain"*. Al igual que el resto de sistemas MER, se detectan palabras genéricas, como *"Finding"* o *"Medical History"* que se marcan como entidad cuando en realidad no lo son.

5.1.3. Discusión

Con respecto a la evaluación del sistema MER de MiNerDoc, en base a los resultados obtenidos (ver Sección 5.1.2) y teniendo en cuenta la limitación del corpus anotado utilizado como Gold Standard, hemos podido obtener algunas conclusiones que nos pueden dar una visión general del rendimiento de nuestro sistema. Proponíamos en este experimento comparar el rendimiento de nuestro sistema MER frente a dos conocidos sistemas de reconocimiento de entidades médicas denominados CliNER y CLAMP. Para evaluar estos sistemas, se procedió a crear una colección de informes de alta que fueron anotados manualmente por un experto en documentación clínica para ser utilizado como *Gold Standard*. El objetivo de este Gold Standard era determinar las entidades médicas que debían ser consideradas como correctamente etiquetas por cada sistema MER evaluado. Este experimento quería demostrar varios aspectos como cuál era el desempeño global del sistema MER de MiNerDoc (basado en MetaMap y UMLS), determinar si dicho enfoque, basado en diccionarios, podría ser competitivo frente a enfoques basados en AA o en reglas o determinar que errores en el reconocimiento de las entidades médicas se han detectado en nuestra propuesta y en el resto de sistemas evaluados. En primer lugar hay que indicar que se trata de un experimento limitado debido principalmente al reducido conjunto de informes anotados que conforman el Gold Standard, esto se debe fundamentalmente a la dificultad de crear un corpus semánticamente anotado en el que se incluyan varios tipos de entidades médicas (enfermedad/procedimiento/tratamiento) y a la dificultad de encontrar anotadores expertos que puedan realizar esta ardua tarea. Actualmente los corpus con anotación semántica de textos clínicos son muy escasos, se focalizan en un único tipo de entidad, (habitualmente la entidad enfermedad) y en su mayoría no son de acceso público. Teniendo en cuenta este difícil marco de evaluación, hemos podido obtener algunas conclusiones que pueden dar una visión del desempeño global de nuestro sistema MER. En cuanto a la métrica precisión, observamos que el sistema MER de MiNerDoc obtiene

los mejores resultados, con un valor de 77.94%, el segundo lugar lo ocupa el sistema CliNer, con un valor del 76.70% y por último se encuentra el sistema CLAMP, con un valor en precisión del 73.85%. Esto quiere decir que para este específico experimento el sistema MER propuesto encontró un mayor número de entidades médicas correctas que las obtenidas por el resto de los sistemas evaluados. Para la métrica Recall, que representa el número de entidades que el sistema MER predijo correctamente entre el número de entidades identificadas en el *Gold Standard*, el sistema que ocupó el primer puesto fue el sistema CLAMP, con un valor de 86.56%, en segundo lugar le sigue nuestro sistema MER con un resultado del 85.48% y por último, el sistema CliNER con un resultado del 84.95%. El sistema MER propuesto obtuvo el mejor resultado en cuanto a la métrica FMeasure, obteniendo un resultado de 81.54%, CliNER logró el segundo mejor resultado obteniendo un valor de 80.61% y el último puesto fue para el sistema CLAMP, con un valor de 79.70%. Por lo tanto, según el análisis de los resultados de esta fase experimental, es posible afirmar que el rendimiento general de nuestro sistema MER (basado en diccionarios) es superior a los sistemas CliNER y CLAMP.

Aunque a simple vista pueda parecer que existe escasa diferencia en el desempeño general del sistema MER de MiNerDoc y el sistema CliNER, en la práctica existen algunas características de especial relevancia en el ámbito médico que diferencian el sistema propuesto del resto. Dos de estas destacadas características son la detección de entidades negadas (no disponible en el sistema CliNER) y la resolución (detección e interpretación) de acrónimos (CliNER y CLAMP detecta acrónimos pero no los interpreta). Ambas ventajas son facilitadas por el sistema MER de MiNerDoc gracias a la utilización de la herramienta MetaMap en nuestro sistema de MT.

En resumen, aunque esta experimentación es algo limitada, hemos podido responder a las preguntas que nos planteábamos al inicio de este experimento, nuestra propuesta, basada en diccionarios, obtiene un buen desempeño global (81.54% en FMeasure) con algunas ventajas sobre el resto de sistemas, como la detección de entidades negadas y

la resolución de acrónimos, pero también hemos detectado algunas deficiencias, que deberán ser corregidas en un futuro, como la detección de entidades incorrectas y la opción de configurar el sistema por el usuario. Si bien es cierto, que la tasa de entidades incorrectas detectadas por el sistema MER propuesto es el más bajo del resto de sistemas (22% de errores frente al 23% de CliNer y 26% de CLAMP), sería necesario en una futura revisión de la herramienta poder disminuir aún más esta tasa de errores.

En resumen hemos comprobado, en base a la experimentación realizada, que el sistema MER de MiNerDoc, basado en diccionarios, obtuvo el mejor resultado en las métricas FMeasure y Precisión, aportando además importantes características esenciales en el área médica, como son la detección de entidades negadas y la resolución de acrónimos.

5.2. Experimentación 2: determinar qué metodología y parametrización mejora el desempeño del sistema CDA

En este segundo experimento realizaremos una evaluación del sistema CDA de MiNerDoc en una doble vía, por un lado, evaluaremos el método de clasificación diagnóstica propuesto en esta tesis doctoral, denominado dCSE (*diagnostic Classification with Semantic Enrichement*) y por otro, determinaremos que parametrización (combinación de distintas técnicas de MT) logra el mejor rendimiento predictivo en la categorización diagnóstica. Para ello, en primer lugar, evaluaremos el rendimiento de la metodología dCSE (ver Sección 3.3.2) frente a la metodología baseline que sigue el modelo convencional de bolsa de palabras sin enriquecimiento semántico ni

terminológico. En segundo lugar, evaluaremos distintos tipos de parametrizaciones utilizando distintas técnicas de MT (métodos *stemming*, eliminación de *stop-words* y diferentes tipos de tokenizaciones) para determinar cuál es la mejor combinación para obtener un mayor rendimiento de la tarea CDA. Estas parametrizaciones serán aplicadas sobre el conjunto de características extraídas de la colección original de informes de alta y producirán ocho datasets diferentes. A continuación, en las siguientes secciones analizaremos la configuración aplicada para llevar a cabo este segundo experimento y los resultados obtenidos.

5.2.1. Configuración experimental

En esta sección detallaremos en primer lugar las principales características del conjunto de datos de partida, formado por la colección de informes de alta procedentes la base de datos MIMIC, que utilizaremos para evaluar el sistema CDA de MiNerDoc. A continuación, para determinar la mejor parametrización, se describirán los diferentes datasets construidos en base a las distintas combinaciones de técnica de MT (*stemming*, no *stemming*, distintos niveles de granularidad, aplicación *stop-words*) y para cada metodología (dCSE versus MetaMap). Por último, detallaremos los distintos métodos MLL y las métricas empleadas en este segundo análisis experimental.

5.2.1.1. Conjunto de datos de partida

Para llevar a cabo la evaluación del sistema de clasificación diagnóstica propuesto, se empleó un conjunto de 1,210 informes de alta de la base de datos MIMIC [55]. Esta base de datos es una de las pocas que permiten un acceso abierto para investigadores del ámbito de la salud (accesible a través de la web de PhysioNet [235]). Contiene datos anonimizados y que incluyen informes con contenido textual clínico procedente de entornos sanitarios reales. Un fragmento de un informe de alta de la colección original

(SemEval 2014-MIMIC II) lo podemos ver en la Figura 5.3 y 5.4. La mayoría de los 1,210 informes seleccionados siguen una estructura similar dividida en varias secciones que se detallan en la Tabla 5.4. Se seleccionaron inicialmente sólo los informes de alta que contenían la evolución completa de la estancia del paciente desde la perspectiva de distintas especialidades médicas, al tratarse en su gran mayoría de pacientes pluripatológicos (sección "*Brief Hospital Course*" o "*Hospital Course*"). Todos los datos personales de los informes se encontraban perfectamente anonimizados gracias a la labor previa realizada en la base de datos MIMIC [55]. Cada informe de alta fue codificado manualmente por un médico experto en documentación clínica del Hospital Universitario Reina Sofía, incluyendo uno o más descriptores de enfermedad MeSH (problema multietiqueta). En este proceso de asignación manual, se consideraron un total de 22 jerarquías de diagnóstico (clases) procedentes de dos categorías MeSH, "*C-Diseases*" y "*F-Psychiatry and Psychology*" (ver Tabla 5.5). Se seleccionaron estas jerarquías diagnósticas por tratarse de los grupos de enfermedad predominantes en la colección MIMIC (pacientes pluripatológicos de una unidad de cuidados intensivos).

Secciones de un informe de alta	
Subject id	Radiology/imaging
Admission date	Physical examination on presentation
Date of birth	Pertinent laboratory data on presentation
Sex	Radiology/imaging
Service	Hospital course by system
Major Surgical or Invasive Procedure	Discharge status
Chief complaint	Discharge diagnoses
History of present illness	Discharge Instructions
Past medical history	Medications on discharge
PAST SURGICAL HISTORY	Discharge diet
Allergies	Discharge activity
Medications on admission	Code status
Social history	Note
FamilyHistory	Dictated by
Physical examination on presentation	Job Number
Pertinent laboratory data on presentation	Signed electronically by

Tabla 5.4. Secciones de un informe de alta tipo original de la colección MIMIC

```
24805 |||| 112 |||| 2028 |||| DISCHARGE_SUMMARY |||| 2012-05-02
00:00:00.0 ||||       ||||       ||||       ||||
Admission Date: [**2012-04-27**]        Discharge Date: [**2012-05-02**]

Date of Birth: [**1812-04-27**]        Sex:  M

Service:

HISTORY OF PRESENT ILLNESS:  The patient is a [**Age over 90 **]-year-old
man with a history of peptic ulcer disease, coronary artery disease,
status post myocardial infarction in [**1996**] as well as 2003, temporal
arteritis, who presented with melenas  and chest pain. The patient
reported melanotic stools times 5 since 4 p.m. on the day prior to
admission. No hematemesis or hematochezia. Stools were loose. The patient
had a history of melena in [**2010-03-29**]. The patient also reported
being lightheaded, fatigued with an increase in his ch set discomfort for
which he was taking sublingual nitroglycerin with relief. On the a.m. of
presentation, the symptoms persisted; the patient contacted his PCP who
sent him to the [**Doctor First Name 5**]. In the [**Doctor First Name
5**], the p was found to have a hematocrit of 22.9 decreased from a
baseline of 32 to 38. He was given IV
Protonix, IV fluids, and transfused the first of 2 units of packed red
blood cells. Gastroenterology was consulted. The patient initially had an
EKG with slight inferior changes while the patient was pain free. The
patient then had an episode of 8/10 substernal chest pain in the
[**Doctor First Name 5**] with 3 to [**Street Address 5847**] changes in
V3 to V4.

PAST MEDICAL HISTORY:  Significant for upper gastrointestinal bleed in
[**2010-03-29**]. An esophagogastroduodenoscopy showed an ulcer in the
pylorus and chronic gastritis, coronary artery disease, status post
myocardial infarction in [**1996**] and 2003, benign prostatic
hypertrophy, history of temporal arteritis, pemphigoid, history of
anemia, history of small bowel volvulus, status post appendectomies,
status post inguinal hernia repair x2, history of colonic polyps, and
sigmoid diverticulosis.

ALLERGIES:  THE PATIENT HAS NO KNOWN DRUG ALLERGIES.

MEDICATIONS:  The patient was on:

1. Celebrex.
2. Aspirin.
3. Prednisone.
4. Atenolol.
5. Imdur.
6. Nitroglycerin p.r.n.

SOCIAL HISTORY:  He is a retired physician, psychiatrist.
Remote tobacco history. Social alcohol use, which is infrequent. Married
with 1 son.

FAMILY HISTORY:  Noncontributory.

PHYSICAL EXAMINATION ON ADMISSION AS FOLLOWS:  VITAL SIGNS:
Vital signs of 98.9 temperature, blood pressure 128/80, pulse 72,
respiratory rate of 13, and oxygen 100% on 3 liters.
GENERAL: The patient appeared comfortable.
```

Figura 5.3.- Fragmento Informe de Alta original (parte I)

HEENT: Examination was unremarkable except for pale conjunctiva, dry mucosa.

LABORATORY DATA: Significant for the hematocrit of 23 as stated above, a potassium of 5.3, a BUN 78. Initial CK was 107 with an MB of 6 and a troponin of 0.02. INR was 1.0.
Urinalysis was unremarkable. As stated above, the patient had 2 EKGs and the second of which showed 2 to [**Street Address 4639**] changes in V3 to V6. Chest x-ray showed no acute cardiopulmonary process, so the patient was admitted to the hospital.

CONCISE SUMMARY OF HOSPITAL COURSE AS FOLLOWS: GI: The patient was felt to likely have another bleeding ulcer as the etiology of his melanotic stools and anemia. The patient had a history of Helicobacter pylori in the past that was treated. The patient was felt to require EGD to evaluate for recurrent infection as well as ongoing bleeding. The patient was initially admitted to the ICU. Gastroenterology was consulted. The patient was taken for EGD on [**04-27**], which showed a deep antral ulcer, no acute bleeding. The ulcer was injected.

The patient was initially continued on IV b.i.d. Protonix.
Hematocrits were followed and the patient was maintained on 2 peripheral IV's at all times, and aspirin was held. The patient has another episode of melanotic stool. On [**2012-04-28**], he was taken for another EGD, at that time which showed the ulcer was not bleeding. As a result, the patient was felt to be stable for discharge to home from a GI perspective with continuation of the b.i.d. Protonix. The patient to follow up for a repeat endoscopy in 8 weeks as an outpatient.

Cardiac: Cardiac enzymes had been significant for elevated troponin on admission. Cardiology was contacted who did not recommend cardiac catheterization or coronary artery bypass graft. The patient initially received heparin and was restarted on aspirin, which was approved by GI as long as the patient had serial hematocrits. The patient was transfused to keep the hematocrit above 30. He was restarted on atenolol.
The patient was also on Imdur for a longer-acting vasodilator effect. The patient had a couple of episodes of further chest pain during the admission but had no further EKG changes.

Pulmonary: The patient had some desaturations to 70's and 80's with ambulation without improvement with oxygen with ambulation, but at this time the patient was completely asymptomatic and the patient's oxygen saturation recovered spontaneously to the high 90's on room air with rest. As a result, this was felt to possibly be not reflective of the patient's pulmonary status, but reflective of some peripheral vascular changes with ambulation. The patient was not felt to need inpatient workup and will follow up with PCP as an outpatient.

Hematology: The patient with acute blood loss anemia, received a total of 4 units of packed red blood cells, had serial hematocrits while on heparin gtt and was transfused to keep the hematocrit above 30.

Musculoskeletal: The patient was restarted on his prednisone for polymyalgia rheumatica and temporal arteritis.

Figura 5.4.- Fragmento Informe de Alta original (parte II)

Clases MeSH	
C01	Bacterial Infections and Mycoses
C02	Virus Diseases
C04	Neoplasms
C05	Musculoskeletal Diseases
C06	Digestive System Diseases
C07	Stomatognathic Diseases
C08	Respiratory Tract Diseases
C10	Nervous System Diseases
C11	Eye Diseases
C12	Male Urogenital Diseases
C13	Female Urogenital Diseases and Pregnancy Complications
C14	Cardiovascular Diseases
C15	Hemic and Lymphatic Diseases
C16	Congenital, Hereditary, and Neonatal Diseases and Abnormal
C17	Skin and Connective Tissue Diseases
C18	Nutritional and Metabolic Diseases
C19	Endocrine System Diseases
C20	Immune System Diseases
C23	Pathological Conditions, Signs and Symptoms
C25	Chemically-Induced Disorders
C26	Wounds and Injuries
F03	Mental Disorders

Tabla 5.5. Jerarquías diagnósticas MeSH (clases) utilizadas en el sistema de clasificación diagnóstica de MiNerDoc

La estadística básica de la colección de informes originales una vez categorizados dentro de las jerarquías diagnósticas MeSH se recoge en las Tablas 5.6 y 5.7. Hemos empleado el paquete *mldr* [221] del software R para llevar a cabo este sumario estadístico de la colección de informes de altas. En la Tabla 5.6 se recogen las medidas cardinalidad y densidad. La *cardinalidad* es el número medio de etiquetas por instancia (Ecuación 5.1) y la *densidad* es una medida que analiza la dispersión de etiquetas (Ecuación 5.2), siendo D un conjunto de datos multietiqueta, L el conjunto total de etiquetas que pueden aparecer en las instancias de D, $|D|$ el número de instancias en D, $|L|$ el número de etiquetas en L e Y_i el subconjunto de L que aparece en la i-ésima instancia de D [221]. Analizando los resultados obtenidos podemos observar que la densidad es baja por lo que existe una dispersión importante de etiquetas, es decir, la colección está desbalanceada ya que no todas las clases se distribuyen por igual.

$$Card = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i| \quad (5.1) \qquad Dens = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i|}{|L|} \quad (5.2)$$

Estadística	
Nº de Instancias	1,210
Nº de Etiquetas	22
Cardinalidad	6.5967
Densidad	0.2998

Tabla 5.6. Estadísticas básicas colección inicial

Label	Nº de Informes	Frecuencia
class-C01	482	0.39834711
class-C02	110	0.09090909
class-C04	400	0.33057851
class-C05	131	0.10826446
class-C06	494	0.40826446
class-C07	44	0.03636364
class-C08	637	0.52644628
class-C10	767	0.63388430
class-C11	71	0.05867769
class-C12	189	0.15619835
class-C13	217	0.17933884
class-C14	850	0.70247934
class-C15	359	0.29669421
class-C16	47	0.03884298
class-C17	157	0.12975207
class-C18	477	0.39421488
class-C19	298	0.24628099
class-C20	168	0.13884298
class-C23	1019	0.84214876
class-C25	351	0.29008264
class-C26	170	0.14049587
class-F03	544	0.44958678

Tabla 5.7. Nº de instancias por clase

Como podemos ver las tres clases con un mayor número de instancias son (ver Figura 5.5) "*C23-Pathological Conditions, Signs and Symptoms*" con 1019 instancias, "*C14-Cardiovascular Diseases*" con 850 instancias y "*C10-Nervous System Diseases"* con 767 instancias. Y las tres clases que tienen una menor frecuencia son: "*C07-Stomatognathic Diseases"* con 44 instancias, "*C16-Congenital, Hereditary, and Neonatal Diseases and Abnormalities*" con 47 instancias y "*C11-EyeDiseases*" con 71 instancias.

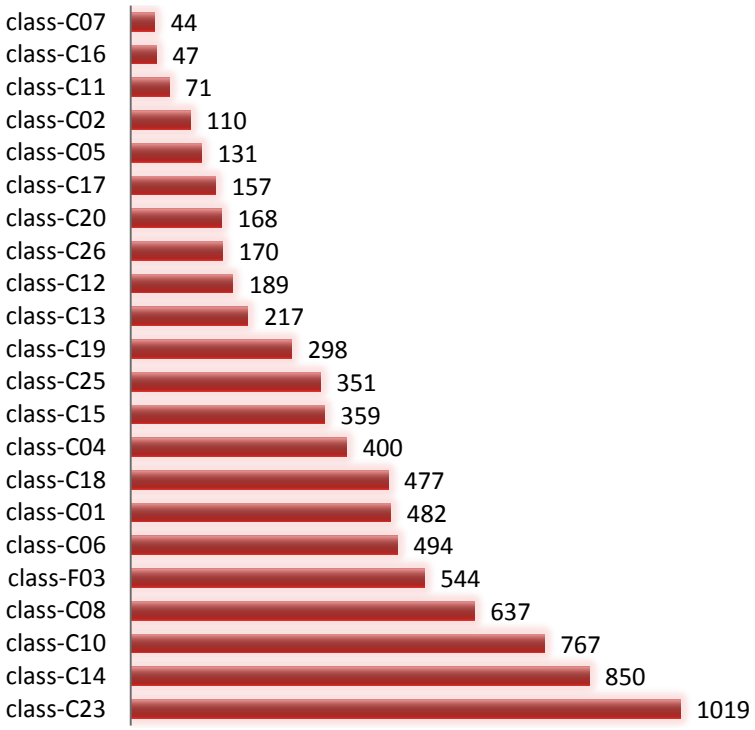


Figura 5.5. Nº de instancias por clase

5.2.1.2. Parametrizaciones

En este segundo experimento distintas parametrizaciones basadas en MT (*stemming*, no *stemming*, distintos niveles de granularidad en la tokenización) han sido aplicadas sobre los conjuntos de características extraídas de la colección original de informes de

de alta de la base de datos MIMIC (1,210 informes), con el objetivo de verificar cuál de ellas ofrece un mayor rendimiento predictivo en la tarea CDA. Cada una de estas combinaciones de parámetros producen ocho conjuntos de datos diferentes (transformación del original), tal y como se muestra en la Tabla 5.8. Se construyeron cuatro conjuntos de datos (denominados datasets MetaMap) aplicando una ingeniería de características con el enriquecimiento semántico proporcionado por la herramienta MetaMap y el metathesaurus UMLS (siguiendo la metodología dCSE vista en la Sección 3.3.2) y cuatro conjuntos de datos (denominados datasets Baseline), siguiendo las mismas parametrizaciones que los datasets anteriores pero construidos sin la aplicación de la herramienta externa MetaMap. Para la selección de los rasgos que conforman los

	DATASETS	Parametrización	Ejemplo de Atributos	Atributos
Datasets dCSE	**dCSE-Param1**	Porter Stemming Stop-Words Tokenización Unigrama	'anticoagul' 'antigen' 'antimicrobi'	2.887
	dCSE-Param2	No Porter Stemming Stop-Words Tokenización Unigrama	'anticoagulant' 'anticoagulation' 'antimicrobial'	3.227
	dCSE-Param3	No Porter Stemming Stop-Words Tokenización Bigramas	'stenosis renal' 'stomach hemorrhage' 'diabetes mellitus'	34.262
	dCSE-Param4	Porter Stemming Stop-Words Tokenización Bigramas	'stenosis ren' 'stomach hemorrhag' 'diabetes mellitu'	33.691
Datasets Baseline	**Base-Param1**	Porter Stemming Stop-Words Tokenización Unigrama	'myocardi' 'mellitu' 'meningit'	3.000
	Base-Param2	No Porter Stemming Stop-Words Tokenización Unigrama	'myocarditis' 'meningitis' 'myocarditis'	3.000
	Base-Param3	No Porter Stemming Stop-Words Tokenización Bigramas	'bilateral metastatic' 'brain metastases' 'breast primary'	35.000
	Base-Param4	Porter Stemming Stop-Words Tokenización Bigramas	'bilateral metastat' 'brain metastas' 'breast primari'	35.000

Tabla 5.8. Datasets generados bajo metodologías dCSE y Baseline (parametrizaciones)

datasets creados bajo el *método baseline* (modelo de referencia) se siguió el modelo convencional *BoW* [127], donde cada informe de alta, después de un preprocesamiento previo de los informes de alta originales ("limpieza" preliminar), fue segmentado en tokens obteniendo una colección de características que formaron los distintos datasets baseline. Esta metodología sólo tiene en cuenta las palabras (tokens) que forman parte de cada informe de alta, no se ha utilizado la herramienta MetaMap ni el metatesauro UMLS, es decir, no se ha empleado ninguna fuente de conocimiento externo para procesar los informes de alta originales y por tanto, no se ha añadido ningún tipo de enriquecimiento semántico a dichas colecciones. Debido al gran volumen de atributos generados bajo el método baseline, fue aplicado un proceso de selección de características para intentar reducir la dimensionalidad de los conjuntos de datos [220]. Las distintas parametrizaciones aplicadas a los conjuntos de características obtenidas bajo la metodología dCSE (denominadas dCSE-Param1, dCSE-Param2, dCSE-Param3 y dCSE-Param4) también han sido aplicadas a los conjuntos de rasgos obtenidos bajo la metodología baseline (Base-Param1, Base-Param2, Base-Param3 y Base-Param4).

5.2.1.3. Métodos multietiqueta y métricas

Se aplicaron los métodos más populares y ampliamente utilizados para resolver la tarea de clasificación multietiqueta [229], considerando los métodos de transformación de problemas, los métodos de adaptación y los métodos ensembles. En cuanto a los métodos de transformación de problemas, se consideraron los siguientes: método *Binary Relevance* (BR) con *Sequential Minimal Optimization* (SMO) como clasificador base (descrito en la literatura como una combinación óptima para este tipo de tareas [57]), *Classifier Chains* (CC) y *Label Powerset* (LP), usando ambos el algoritmo J48 como clasificador base (algoritmo por defecto para estos métodos). En cuanto a los métodos de adaptación, se emplearon los siguientes: *AdaBoost.MH, instance-based learning by*

logistic regression (IBLR) y *multi-label k-nearest neighbors* (ML-Knn). Finalmente, se consideran los siguientes métodos ensembles: *ensemble classifier chain* (ECC), *ensemble of pruned set* (EPS), *hierarchy of multi-label classifiers* (HOMER), *multi-label stacking* (MLS) y *ramdom klabelsets* (RakEL). Se realizó una validación cruzada estratificada en 5-folds para cada dataset, utilizando el procedimiento iterativo de estratificación [219]. Con el particionado estratificado se intentará mitigar el riesgo que generan las colecciones desbalanceadas, como son las empleadas en esta experimentación, ya que se consigue que en cada partición haya igual número de instancias de cada clase representada. Adicionalmente, los métodos multietiqueta considerados en esta fase experimental fueron ejecutados con 10 semillas distintas.

Por último, las métricas de evaluación de los clasificadores multietiqueta pueden englobarse en dos grandes grupos [229]: *métricas basadas en ejemplos y métricas basadas en etiquetas*. Con el primer grupo de métricas se toma como entrada la respuesta del clasificador para cada instancia y posteriormente se promedia entre el número de instancias totales. Las métricas basadas en etiquetas evalúan cada etiqueta por separado siendo posteriormente promediadas para obtener un único valor, para promediar se utilizan dos aproximaciones, macro y micro. Se han utilizado 4 métricas de evaluación, dos métricas basadas en ejemplos como Hamming Loss y FMeasure (FMeasure_{ex}) y dos basadas en etiquetas, FMeasure micro (FMeasure_{mic}) y FMeasure macro (FMeasure_{mac}). En la tabla 5.9 se definen formalmente las métricas de evaluación citadas anteriormente, donde t es el número de instancias del conjunto test, q es el número de etiquetas, Yi representa el número de etiquetas reales, Zi representa el número de etiquetas predichas, Δ representa la diferencia simétrica entre dos conjuntos de etiquetas, tp indica el número de patrones positivos clasificados correctamente, fp indica el número de falsos positivos, es decir, los patrones negativos que han sido correctamente asignados, y por último, fn que indica el número de falsos negativos, es decir, los patrones que no han sido asignados.

Métricas de evaluación multietiqueta	
Métricas basadas en ejemplos	
Hamming Loss ↓	$\frac{1}{t}\sum_{i=1}^{t}\frac{1}{q}\lvert Z_i \triangle Y_i\rvert$
FMeasure$_{ex}$ ↑	$\frac{1}{t}\sum_{i=1}^{t}\frac{2\lvert Z_i \bigcap Y_i\rvert}{\lvert Z_i\rvert+\lvert Y_i\rvert}$
Métricas basadas en etiquetas	
FMeasure$_{mac}$ ↑	$\frac{1}{q}\sum_{i=1}^{q}\frac{2\cdot tp_i}{2\cdot tp_i+fn_i+fp_i}$
FMeasure$_{mic}$ ↑	$\frac{\sum_{i=1}^{q}2\cdot tp_i}{\sum_{i=1}^{q}2\cdot tp_i+\sum_{i=1}^{q}fn_i+\sum_{i=1}^{q}fp_i}$

Tabla 5.9. Métricas de evaluación multietiqueta

5.2.2. Resultados

La Tabla 5.10 muestra los resultados globales para cada métrica y metodología de evaluación (los mejores resultados se muestran en negrita). Realizando un análisis inicial de los resultados y tomando en consideración los mejores resultados obtenidos para cada metodología (dCSE o Baseline), podemos observar que el mejor valor para la métrica ***Hamming Loss*** (métrica a minimizar) fue ***0.0879*** bajo la metodología dCSE con la parametrización 2 (dCSE-Param2) y ***0.1973*** para la metodología Baseline con la parametrización 2 y 4 (Base-Param2 y Base-Param4). Para la métrica ***FMeasure$_{ex}$*** (métrica a maximizar), dCSE obtiene ***0.8424*** mientras el método de referencia obtiene ***0.6138.*** Con respecto a la métrica ***FMeasure$_{mic}$*** (métrica a maximizar), dCSE obtiene un valor de ***0.8483*** y baseline ***0.6347.*** Para la métrica ***FMeasure$_{mac}$*** (métrica a maximizar),

dCSE obtuvo ***0.7122*** y baseline ***0.4176***. Como observamos en base a estos resultados analizados inicialmente, *la metodología dCSE obtiene mejores resultados para cada una de las cuatro métricas de evaluación frente a la metodología baseline.*

El siguiente paso de este análisis inicial se centró en conocer si los valores promedio para todos los algoritmos y todas las parametrizaciones realizadas bajo la metodología dCSE superaron a los valores promedio obtenidos bajo la metodología baseline. La Figura 5.6 muestra el rendimiento global de todas las parametrizaciones de cada metodología, dCSE y Baseline, para cada métrica y método MLL. Como podemos observar, el rendimiento de la metodología propuesta en esta investigación (dCSE) es superior a la metodología baseline para cualquiera de las 4 métricas. Para apoyar y validar los análisis previos, se llevó a cabo un análisis estadístico para determinar si existían diferencias estadísticamente significativas entre ambas metodologías (dCSE y Baseline) bajo las mismas parametrizaciones. Para llevar a cabo esta comparación, se aplicó el test de Wilcoxon [218]. A través de este test no paramétrico se determinará si se cumple la hipótesis nula, es decir, si no hay diferencias significativas entre el rendimiento de los dos métodos analizados, o si se rechaza la hipótesis nula, es decir, sí existen diferencias significativas entre los dos métodos. El resultado se recoge en la Tabla 5.11, como se observa, considerando un nivel de confianza del 95% (*p-value* menor que 0.05), el *método dCSE supera al método baseline para todas las métricas analizadas* (Hamming loss, $FMeasure_{ex}$, $Fmeasure_{mic}$ y $FMeasure_{mac}$).

Este análisis inicial revela que gracias a la información proporcionada por UMLS y la herramienta MetaMap ha sido posible realizar una interpretación semántica de los informes de alta, hecho que añade enriquecimiento a las características que conforman los datasets construidos bajo la metodología propuesta, incorporando peculiaridades como la detección de negaciones, interpretación de acrónimos, desambiguación de términos y otros elementos de interés en el ámbito clínico que hacen que la metodología dCSE mejore el rendimiento de la tarea de clasificación diagnóstica multietiqueta.

Métricas	Datasets	BR	CC	LP	AdaBoost.MH	IBLR	MLkNN	ECC	EPS	HOMER	MLS	RakEL
Hamming Loss	dCSE-Param1	0.1092	0.1085	0.2400	0.2687	0.1879	0.1900	0.0934	0.2182	0.1258	0.1109	0.0889
	dCSE-Param2	0.1093	0.1104	0.2485	0.2687	0.1873	0.1908	0.0965	0.2209	0.1293	0.1141	**0.0879**
	dCSE-Param3	0.1196	0.1191	0.2563	0.2688	0.2048	0.2115	0.1030	0.2227	0.1347	0.1232	0.1208
	dCSE-Param4	0.1193	0.1190	0.2523	0.2688	0.2057	0.2103	0.1013	0.2211	0.1329	0.1189	0.1192
	Base-Param1	0.2328	0.2346	0.2899	0.2686	0.2240	0.2287	0.2024	0.2358	0.2524	0.2334	0.2054
	Base-Param2	0.2251	0.2294	0.2925	0.2686	0.2276	0.2287	**0.1973**	0.2364	0.2497	0.2334	0.2031
	Base-Param3	0.2407	0.2412	0.2970	0.2686	0.2312	0.2317	0.2027	0.2368	0.2579	0.2403	0.2075
	Base-Param4	0.2353	0.2346	0.2975	0.2686	0.2319	0.2338	**0.1973**	0.2361	0.2565	0.2369	0.2077
$FMeasure_{ex}$	dCSE-Param1	0.7988	0.8000	0.5804	0.2333	0.6238	0.6141	0.8212	0.5177	0.7708	0.7959	0.8413
	dCSE-Param2	0.7961	0.7934	0.5641	0.2333	0.6192	0.6117	0.8121	0.5113	0.7643	0.7871	**0.8424**
	dCSE-Param3	0.7728	0.7718	0.5514	0.2333	0.5744	0.5662	0.7992	0.5078	0.7539	0.7671	0.7665
	dCSE-Param4	0.7756	0.7764	0.5607	0.2333	0.5742	0.5621	0.8039	0.5136	0.7594	0.7792	0.7702
	Base-Param1	0.5590	0.5531	0.4902	0.2337	0.5428	0.5295	0.5941	0.4685	0.5653	0.5607	0.6106
	Base-Param2	0.5728	0.5640	0.4879	0.2337	0.5356	0.5267	0.6064	0.4668	0.5700	0.5624	**0.6138**
	Base-Param3	0.5477	0.5453	0.4723	0.2337	0.5287	0.5072	0.5922	0.4661	0.5593	0.5498	0.6065
	Base-Param4	0.5630	0.5628	0.4728	0.2337	0.5200	0.5007	0.6080	0.4680	0.5527	0.5613	0.6132
$FMeasure_{mic}$	dCSE-Param1	0.8104	0.8113	0.5924	0.2218	0.6459	0.6313	0.8342	0.5134	0.7856	0.8071	0.8474
	dCSE-Param2	0.8083	0.8067	0.5782	0.2219	0.6442	0.6291	0.8267	0.5056	0.7785	0.7996	**0.8483**
	dCSE-Param3	0.7886	0.7889	0.5645	0.2217	0.6020	0.5811	0.8145	0.5039	0.7706	0.7828	0.7849
	dCSE-Param4	0.7888	0.7901	0.5738	0.2217	0.6011	0.5780	0.8182	0.5103	0.7738	0.7911	0.7863
	Base-Param1	0.5868	0.5822	0.5122	0.2218	0.5647	0.5404	0.6196	0.4645	0.5866	0.5880	0.6303
	Base-Param2	0.6009	0.5932	0.5096	0.2218	0.5533	0.5357	0.6308	0.4616	0.5898	0.5887	**0.6347**
	Base-Param3	0.5759	0.5734	0.4932	0.2218	0.5447	0.5084	0.6169	0.4610	0.5781	0.5766	0.6260
	Base-Param4	0.5842	0.5853	0.4934	0.2218	0.5327	0.5026	0.6282	0.4631	0.5705	0.5836	0.6295
$Fmeasure_{mac}$	dCSE-Param1	0.6647	0.6620	0.4501	0.0416	0.3988	0.3641	0.6608	0.2305	0.6371	0.6558	**0.7122**
	dCSE-Param2	0.6590	0.6622	0.4369	0.0416	0.3996	0.3618	0.6526	0.2176	0.6189	0.6494	0.7082
	dCSE-Param3	0.6362	0.6378	0.4251	0.0416	0.3483	0.2914	0.6459	0.2126	0.6191	0.6267	0.6306
	dCSE-Param4	0.6393	0.6413	0.4309	0.0416	0.3489	0.2930	0.6482	0.2195	0.6195	0.6379	0.6332
	Base-Param1	0.3916	0.3886	0.3551	0.0416	0.2710	0.2256	0.3757	0.1509	0.4003	0.3950	0.4072
	Base-Param2	0.4086	0.4046	0.3502	0.0416	0.2581	0.2134	0.3905	0.1531	0.4058	0.4008	**0.4176**
	Base-Param3	0.3835	0.3822	0.3286	0.0416	0.2350	0.1663	0.3682	0.1445	0.3938	0.3835	0.4032
	Base-Param4	0.3944	0.3962	0.3264	0.0416	0.2204	0.1668	0.3765	0.1455	0.4004	0.3951	0.4011

Tabla 5.10. Resultados promedios obtenidos por cada método MLL para cada métrica y cada dataset (parametrizaciones descritas en Tabla 5.7). Los mejores resultados para cada métrica y métodología se marcan en negrita.

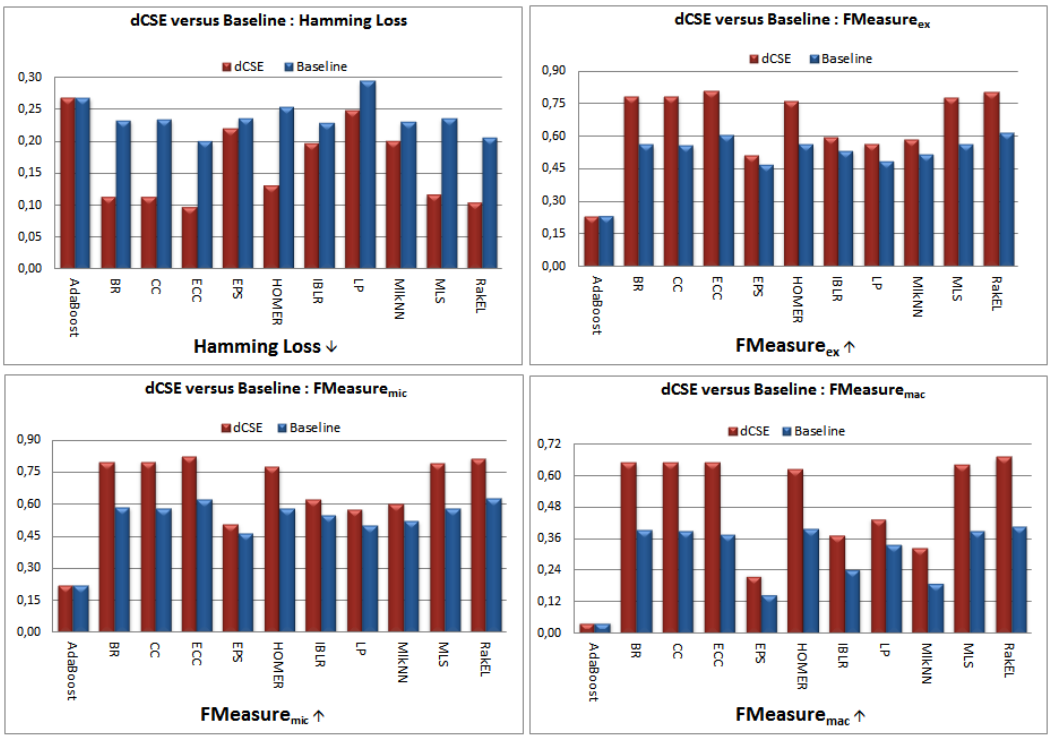


Figura 5.6. Análisis comparativo de la metodología dCSE y Baseline para cada métrica y método MLL.

Después de mostrar los resultados en relación a la evaluación de la metodología, nos centraremos en presentar los resultados obtenidos en relación a la parametrización (ver Tabla 5.7) que permite mejorar el desempeño de la tarea de codificación diagnóstica. Basándonos nuevamente en los valores de la tabla global de resultados (Tabla 5.10), realizamos, en primer lugar, el ranking promedio para cada parametrización (considerando todos los métodos MLL) y cada métrica analizada (ver Tabla 5.12). En base a estos resultados, observamos que la parametrización que obtiene un mejor resultado para las cuatro métricas analizadas es la parametrización *"dCSE-Param1",* es decir, aquella construida bajo la metodología dCSE aplicando la técnica stemming, eliminación de stop-words y una tokenización basada en unigramas. El peor resultado obtenido es para la parametrización *"Base-Param3",* es decir, la parametrización llevada

dCSE-Param1 versus Base-Param1			
Métrica	R^+(dCSE)	R^-(Baseline)	p-value
Hamming loss	65.0	1.0	**0.0022**
FMeasure$_{ex}$	65.0	1.0	**0.0022**
Fmeasure$_{mic}$	65.0	1.0	**0.0022**
FMeasure$_{mac}$	66.0	0.0	**0.0017**

dCSE-Param2 versus Base-Param2			
Métrica	R^+(dCSE)	R^-(Baseline)	p-value
Hamming loss	65.0	1.0	**0.0022**
FMeasure$_{ex}$	65.0	1.0	**0.0022**
Fmeasure$_{mic}$	66.0	0.0	**0.0017**
FMeasure$_{mac}$	66.0	0.0	**0.0017**

dCSE-Param3 versus Base-Param3			
Métrica	R^+(dCSE)	R^-(Baseline)	p-value
Hamming loss	65.0	1.0	**0.0022**
FMeasure$_{ex}$	65.0	1.0	**0.0022**
Fmeasure$_{mic}$	65.0	1.0	**0.0022**
FMeasure$_{mac}$	65.0	1.0	**0.0022**

dCSE-Param4 versus Base-Param4			
Métrica	R^+(dCSE)	R^-(Baseline)	p-value
Hamming loss	65.0	1.0	**0.0022**
FMeasure$_{ex}$	65.0	1.0	**0.0022**
Fmeasure$_{mic}$	65.0	1.0	**0.0022**
FMeasure$_{mac}$	65.0	1.0	**0.0022**

Tabla 5.11.- Resultados Test Wilcoxon (mejor método dCSE o Baseline)

R^+: suma los rangos positivos donde el método dCSE es superior al método Baseline.

R^-: suma los rangos negativos donde el método Baseline es superior al método dCSE.

Métricas	dCSE-Param1	dCSE-Param2	dCSE-Param3	dCSE-Param4	Base-Param1	Base-Param2	Base-Param3	Base-Param4
Hamming Loss	**1.59**	2.11	4.23	3.50	5.41	5.23	7.14	6.77
FMeasure$_{ex}$	**1.50**	2.32	4.14	3.50	5.77	5.32	7.23	6.23
Fmeasure$_{mic}$	**1.55**	1.91	4.14	3.50	5.59	5.32	7.32	6.68
Fmeasure$_{mac}$	**1.18**	2.09	4.23	3.23	6.05	5.23	7.41	6.59
Meta-Ranking	**1.45**	2.11	4.18	3.43	5.70	5.27	7.27	6.57

Tabla 5.12. Ranking Promedio para cada parametrización (dCSE y Baseline) considerando todos los métodos MLL. Los mejores resultados están marcados en negrita.

Dataset construidos bajo método dCSE: dCSE-Param1 dCSE-Param2, dCSE-Param3 y dCSE-Param4.

Dataset construidos bajo método Baseline: Base-Param1, Base-Param2, Base-Param3 y Base-Param4.

En la última fila se muestra el ranking promedio resultante del test de Friedman.

a cabo bajo la metodología baseline, y por tanto, la construida sin la utilización de fuentes externas de conocimiento. Esta parametrización se ha construido utilizando una tokenización basada en bigramas y donde no se ha aplicado la técnica *stemmin*g. Basándonos en los valores meta-ranking podemos obtener una visualización del ranking de parametrizaciones (ver Figura 5.7) que refleja cual ha sido la mejor o peor parametrización para llevar a cabo la tarea de codificación automática propuesta en esta tesis. Podemos comprobar inicialmente que los resultados de la clasificación para las parametrizaciones llevadas a cabo bajo la metodología dCSE son superiores a los resultados obtenidos bajo la metodología baseline.

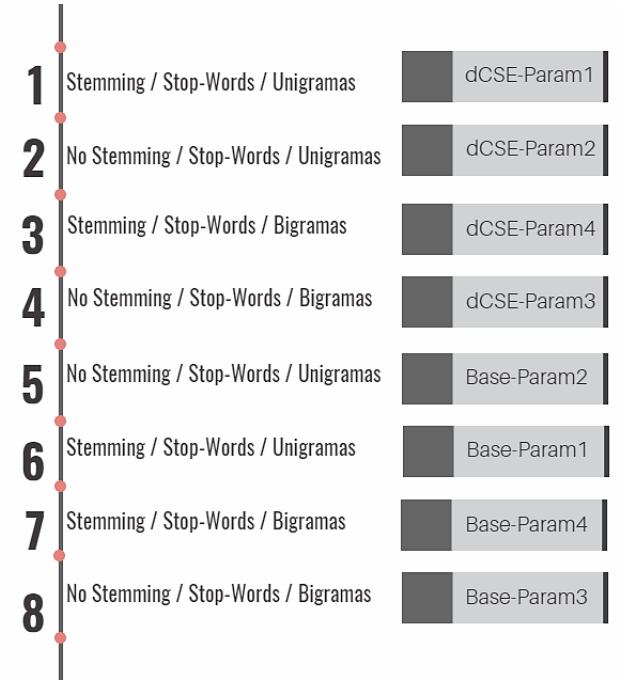


Figura 5.7.- Ranking de parametrizaciones (mejor desempeño a peor desempeño)
Parametrizaciones construidas bajo método dCSE: dCSE-Param1, dCSE-Param2, dCSE-Param3 y dCSE-Param4. Parametrizacionest construidas bajo método Baseline: Base-Param1, Base-Param2, Base-Param3 y Base-Param4.

Para llevar a cabo un análisis más detallado en relación a las parametrizaciones, se realizó un análisis estadístico en dos pasos para comprobar si existían diferencias estadísticamente significativas. En un primer paso se realizará el test de ***Friedman*** [218] para poder rechazar la hipótesis nula, es decir, para poder rechazar la idea de que no hay diferencias entre las distintas parametrizaciones llevadas a cabo para formar el vocabulario de cada dataset. Si el test de Friedman confirma que existen diferencias significativas, realizaremos en un segundo paso un post-hoc test, ***test de Shaffer*** [213], para poder identificar las diferencias existentes entre las distintas parametrizaciones. Al realizar el test de Friedman se observó que considerando una confianza del 95% (p-value menor de 0.05), se rechaza la hipótesis nula para todas las métricas de evaluación, tal y como podemos comprobar en la Tabla 5.13. Por tanto, se confirma que ***existen diferencias estadísticamente significativas entre las distintas parametrizaciones llevadas a cabo para la construcción de los datasets.*** Una vez confirmado que existen diferencias significativas aplicaremos un segundo test para determinar si existen diferencias significativas entra las distintas parametrizaciones. Para ello, aplicaremos el test de Shaffer que permitirá realizar una comparativa múltiple (todos contra todos).

En la Figura 5.8 y en la Tabla 5.14 se muestran los resultados del post-hoc test de Shaffer para cada métrica de evaluación analizada. Considerando un nivel de confianza del 95%, observamos que *la parametrización dCSE-Param1, construida bajo la metodología dCSE,*

Resultados test Friedman (parametrizaciones)	
Métrica	p-value
Hamming loss	4.82E-09
FMeasure$_{ex}$	5.75E-10
Fmeasure$_{mic}$	4.14E-11
FMeasure$_{mac}$	1.10E-08

Tabla 5.13.- Test de Friedman para determinar si existen diferencias estadísticamente significativas entre las distintas parametrizaciones construidas en esta fase experimental.

tiene diferencias significativas con todas las parametrizaciones construidas bajo la metodología Baseline, y como hemos podido observar, teniendo en cuenta el ranking promedio obtenido en el test de Friedman, *es la parametrización más adecuada para llevar a cabo la tarea de clasificación diagnóstica multietiqueta*. Sin embargo, el test de Shaffer no detectó diferencias significativas entre las distintas parametrizaciones construidas bajo el método dCSE (dCSE-Param1, dCSE-Param2, dCSE-Param3 y dCSE-Param4), ni entre las construidas bajo el método Baseline (Base-Param1, Base-Param2, Base-Param3 y Base-Param4), aunque si detectó diferencias entre los *dataset* construidos bajo una metodología y otra.

La principal conclusión de este segundo análisis es que se confirma que todas las parametrizaciones construidas bajo la metodología dCSE hacen aumentar el rendimiento del sistema de clasificación diagnóstico incluido en la aplicación MiNerDoc (comparadas con aquellas parametrizaciones construidas bajo la metodología Baseline). Aunque no se obtuvieron diferencias estadísticamente significativas entre las parametrizaciones construidas bajo la metodología dCSE, se observa que la parametrización "*dCSE-Param1*" es la que mejor desempeño arroja en las cuatro métricas analizadas (ver Figura 5.8). Gracias a la selección de atributos relevantes centrados en el diagnóstico y la eliminación de características redundantes, se evita el ruido innecesario en las colecciones y aumenta la calidad de los *datasets* dCSE, lo que hace mejorar el rendimiento del clasificador.

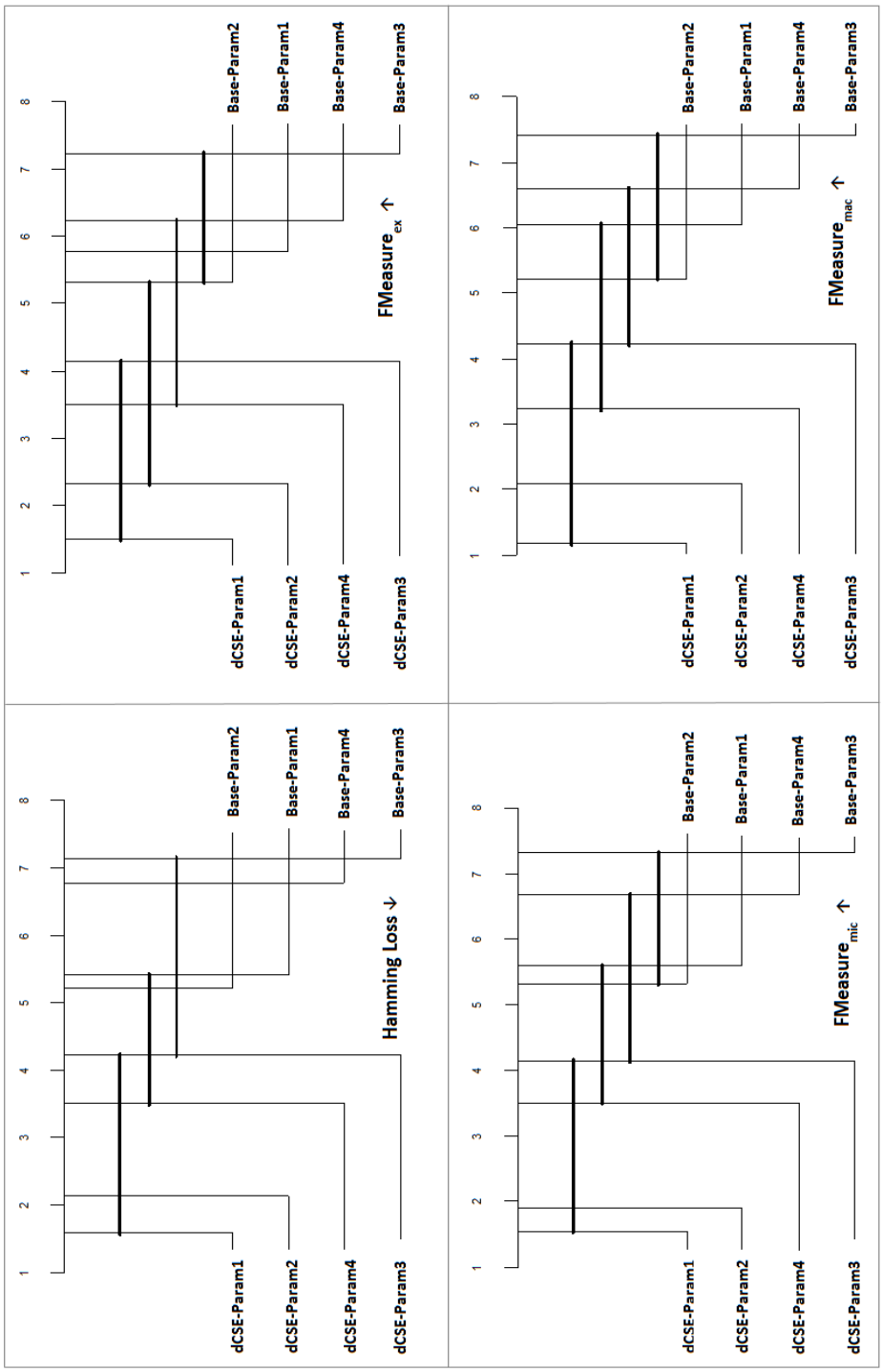


Figura 5.8. Comparación multiple de parametrizaciones Test de Shaffer (nivel de confianza 95%).

	Parametrización	dCSE-Param1	dCSE-Param2	dCSE-Param3	dCSE-Param4	Base-Param1	Base-Param2	Base-Param3	Base-Param4
Hamming Loss	dCSE-Param1	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	dCSE-Param2	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	dCSE-Param3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	dCSE-Param4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
	Base-Param1	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Base-Param2	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Base-Param3	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE
	Base-Param4	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE
FMeasure ex	dCSE-Param1	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	dCSE-Param2	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE	**TRUE**	**TRUE**
	dCSE-Param3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE
	dCSE-Param4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE
	Base-Param1	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Base-Param2	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Base-Param3	**TRUE**	**TRUE**	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE
	Base-Param4	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FMeasure mic	dCSE-Param1	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	dCSE-Param2	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	dCSE-Param3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE
	dCSE-Param4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
	Base-Param1	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Base-Param2	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Base-Param3	**TRUE**	**TRUE**	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE
	Base-Param4	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE
FMeasure mac	dCSE-Param1	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	dCSE-Param2	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	dCSE-Param3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE
	dCSE-Param4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
	Base-Param1	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Base-Param2	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Base-Param3	**TRUE**	**TRUE**	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE
	Base-Param4	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE

Tabla 5.14.- Comparativa múltiple de todas las parametrizaciones mediante test de Shaffer.
True: existen diferencias significativas / False: no existen diferencias significativas
Parametrizaciones bajo método dCSE (propuesto): dCSE-Param1 a dCSE-Param4.
Parametrizaciones bajo método Baseline: Base-Param1 a Base-Param4

5.2.3. Discusión

De los experimentos realizados para evaluar la metodología (dCSE versus MetaMap) y la mejor parametrización para llevar a cabo la tarea CDA en MiNerDoc hemos obtenido varias conclusiones. En un primer análisis, se ha verificado, a través de varios test estadísticos, que la metodología dCSE (metodología propuesta para mejorar la clasificación diagnóstica) es superior en rendimiento al método de referencia para todas las métricas de evaluación. Aunque la base de las dos metodologías (dCSE y Baseline) parten de un modelo de representación simple y ampliamente extendido en el ámbito de la categorización de textos, como es el modelo *BoW*, los procesos seguidos para incorporar conocimiento externo a los diferentes diccionarios creados, aportan un enriquecimiento semántico y terminológico en las colecciones de datos que hace aumentar la efectividad en la clasificación. Gracias a la utilización de MetaMap y el metatesauro UMLS se consigue que el rendimiento predictivo de la tarea de clasificación diagnóstica multietiqueta mejore significativamente. La metodología dCSE ha permitido aumentar la calidad de las características que forman las colecciones de datos y la consecuencia son los buenos resultados obtenidos en este análisis experimental. Gracias a la información que aporta UMLS y la herramienta MetaMap ha sido posible realizar una interpretación semántica de los informes clínicos, incorporando detección de negaciones, interpretación de acrónimos, desambiguación de términos y otras características representativas que han aportado un enriquecimiento terminológico y semántico en los dataset construidos bajo la metodología propuesta (dCSE). Hemos podido comprobar cómo los procesos llevados a cabo con la metodología dCSE han aportado eficacia y un mejor rendimiento en la tarea de clasificación diagnóstica multietiqueta. Por tanto, podemos afirmar que el enriquecimiento semántico proporcionado por la utilización del metatesaurus UMLS, a través de la herramienta Metamap, aporta una mejora importante en el desempeño de la categorización diagnóstica de informes de alta. En base a los resultados recogidos en la Tabla 5.10, y

tomando en consideración los mejores resultados obtenidos para cada metodología (dCSE o Baseline), podemos observar que ***la metodología dCSE aportaría una mejora del 55% (porcentaje de incremento) para la métrica Hamming Loss***. Para la métrica ***FMeasure$_{ex}$ el método propuesto aportaría una mejora del 37%***. Con respecto a la métrica ***FMeasure$_{mic}$ el método propuesto consigue una mejora del 34%***. Para la métrica ***FMeasure$_{mac}$, el método propuesto mejora al método baseline en un 71%.***

En base a varios análisis estadísticos hemos demostrado que el rendimiento predictivo en la tarea clasificación diagnóstica bajo la metodología dCSE es superior a la metodología baseline, gracias al enriquecimiento semántico que aportan las fuentes externas de conocimiento empleadas en MiNerDoc.

El segundo análisis realizado (parametrizaciones) para evaluar el sistema CDA de MiNerDoc ha permitido reforzar las conclusiones obtenidas en el primer análisis. En este análisis, hemos demostrado que cualquier conjunto de características construidas bajo la metodología propuesta, dCSE, con las distintas parametrizaciones evaluadas (tokenización basada en unigramas, tokenización basada en bigramas, aplicación de la técnica *stemmin*g, etc), obtuvieron un mejor rendimiento en la clasificación diagnóstica comparadas con las parametrizaciones construidas bajo la metodología baseline. Los resultados de este segundo análisis (ver Tabla 5.11) nos han demostrado que la parametrización que ha obtenido un mejor resultado es la *dCSE-Param1*, es decir, la parametrización construida con la metodología dCSE, con eliminación de stop-word, tokenización basada en unigramas y con aplicación de la técnica *stemming* (Porter). Como observamos en la Tabla 5.11, las 4 parametrizaciones llevadas a cabo bajo el método dCSE (dCSE-Param1, dCSE-Param2, dCSE-Param3 y dCSE-Param4) arrojan mejores resultados que las parametrizaciones realizadas bajo el método Baseline (Base-Param1, Base-Param2, Base-Param3 y Base-Param4). La parametrización de mejor a peor valor de ranking promedio, según test de Friedman, fue la siguiente: *1ª*) dCSE-

Param1, *2º*) dCSE-Param2, *3º*) dCSE-Param4, *4º*) dCSE-Param3, *5º*) Base-Param2, 6º) Base-Param1, 7º) Base-Param4 y 8º) Base-Param3. Si bien el mejor valor del ranking es para la parametrización dCSE-Param1, la peor parametrización es para la Base-Param4, construida bajo el método Baseline, es decir sin incorporación de enriquecimiento semántico. Para seleccionar los rasgos de este dataset (Base-Param4) se había aplicado la eliminación de stop-words, una tokenización basada en bigramas (pares de palabras) y no se aplicó la técnica de *stemmin*g. Al realizar el test de Friedman (ver Tabla 5.12), considerando una confianza del 95%, pudimos comprobar que se rechazaba la hipótesis nula y por tanto, se confirmó que ***existían diferencias significativas entre las distintas parametrizaciones llevadas a cabo para las construcción de los distintos datasets.*** Gracias al *post-hoc test de Shaffer* se pudo observar las diferencias entre las distintas parametrizaciones, ya que se realizó una comparativa múltiple de todos contra todos (ver Figura 5.6). En la Tabla 5.13 pudimos comprobar las diferencias significativas y no significativas entre las distintas parametrizaciones construidas bajo las dos metodologías analizadas. La parametrización dCSE-Param1, construida bajo la metodología dCSE, obtuvo diferencias significativas (para todas las métricas) con todas las parametrizaciones llevadas a cabo bajo la metodología Baseline , y como hemos podido observar, teniendo en cuenta el ranking promedio obtenido en el test de Friedman, es la parametrización más adecuada para llevar a cabo la tarea de clasificación diagnóstica multietiqueta. La parametrización dCSE-Param2 tiene diferencias significativas con todas las parametrizaciones construidas bajo la metodología Baseline para 3 de las 4 métricas analizadas (Hamming Loss, $FMeasure_{mic}$ y $FMeasure_{mac}$). Aunque los mejores resultados en todas las métricas se han dado para las parametrizaciones del método dCSE, cabe destacar que el test de Shaffer no detectó diferencias significativas entre ninguna de las parametrizaciones construidas bajo el método dCSE, es decir, no hubo diferencias significativas entre las parametrizaciones dCSE-Param1, dCSE-Param2, dCSE-Param3 y dCSE-Param4 entre sí. De igual forma ocurre con las parametrizaciones construidas bajo el método Baseline, Base-Param1, Base-Param2, Base-Param3 y Base-Param4, no

existen diferencias significativas entre ellas. En general, se observa que en cuanto a la aplicación de los diferentes tipos de tokenizaciones (unigrama o bigrama) los mejores resultados se dan en aquellas parametrizaciones donde se aplicó la tokenización basada en unigramas (tanto para la metodología dCSE como para la metodología Baseline). Con respecto a la utilización de la tokenización basada en bigramas, hace que disminuya el rendimiento del clasificador, principalmente motivado por el crecimiento del tamaño de los atributos que conlleva un aumento del ruido en las colecciones de datos. La utilización de la tokenización basada en unigramas reduce la dimensionalidad del vocabulario de los datasets y por tanto, la efectividad del clasificador aumenta. En cuanto a la aplicación de la técnica *stemmin*g, también observamos que su utilización mejoró los resultados en la tarea CDA para ambas metodologías. En nuestra experimentación, bajo la metodología dCSE (la aplicada en MiNerDoc) se ha obtenido un beneficio en el rendimiento del clasificador cuando se aplicó la técnica *stemming*, posiblemente debido a que las colecciones construidas bajo la metodología propuesta parten de unos rasgos de gran calidad y centrados en la categoría diagnóstica, con un ruido casi nulo, por lo que la reducción a la raíz de estos términos difícilmente pueden ir asociados con distorsiones o alteraciones del significado de los términos que perjudiquen el rendimiento predictivo del clasificador.

En resumen, los resultados prometedores obtenidos bajo las parametrizaciones dCSE son la consecuencia de la aplicación de los múltiples procesos llevados a cabo para aportar enriquecimiento semántico a las características que conforman los datasets construidos bajo la metodología dCSE.

La parametrización empleada en MiNerDoc (dCSE-Param1) es superior al resto de parametrizaciones evaluadas para todas las métricas analizadas. Esta parametrización se ha construido aplicando la técnica stemming, la tokenización basada en unigramas y la eliminación de stop-words. Las cuatro parametrizaciones construidas bajo la metodología propuesta (dCSE) han obtenido mejores resultados que las cuatro parametrizaciones construidas bajo la metodología Baseline.

En general, la principal conclusión de este segundo experimento ha sido que la aplicación de la metodología dCSE y la parametrización adecuada (dCSE-Param1) en nuestro sistema MiNerDoc, han permitido transformar un conjunto de palabras, provenientes de informes de alta médica, en una conceptualización de términos focalizados en el diagnóstico. Este hecho ha permitido incorporar, a las colecciones textuales, rasgos de calidad (eliminando características redundantes y carentes de valor) que han incrementado el rendimiento predictivo del sistema MiNerDoc para la tarea CAD.

5.3. Experimentación 3: Determinar qué método de aprendizaje multietiqueta ofrece un mejor resultado en la tarea CDA

El objetivo principal de este tercer experimento es evaluar el rendimiento del modelo de clasificación multietiqueta elegido en MiNerDoc para llevar a cabo la tarea de predicción diagnóstica, para ello, se compararon los métodos más ampliamente utilizados en el estado del arte para resolver la tarea de clasificación multietiqueta. Se determinará qué método es el óptimo (en cualquiera de las parametrizaciones) para mejorar el rendimiento de la tarea CDA.

5.3.1. Configuración experimental

Se compararon 11 métodos multietiqueta diferentes considerando los tres grupos existentes [229], métodos de transformación de problemas, métodos de adaptación de algoritmos y ensembles de clasificadores multietiqueta. Se tuvieron en cuenta los siguientes métodos de transformación de problemas: método Binary Relevance (BR) con Sequential Minimal Optimization (SMO) como clasificador base (descrito en la literatura

como una combinación óptima para este tipo de tareas [57]), Classifier Chains (CC) y Label Powerset (LP), usando ambos el algoritmo J48 como clasificador base (algoritmo por defecto para estos métodos). En cuanto a los métodos de adaptación, se tuvieron en cuenta los siguientes: AdaBoost.MH, instance-based learning by logistic regression (IBLR) y multi-label k-nearest neighbors (ML-Knn). Se emplearon los siguientes métodos ensembles: ensemble classifier chain (ECC), ensemble of pruned set (EPS), hierarchy of multi-label classifiers (HOMER), multi-label stacking (MLS) y ramdom klabelsets (RakEL). Se realizó una validación cruzada estratificada en 5-folds para cada dataset, utilizando el procedimiento iterativo de estratificación [219]. Los métodos multietiqueta considerados en esta fase experimental fueron ejecutados con 10 semillas distintas.

Tal y como recogimos en la Sección 5.2.1.3 (Tabla 5.9), se han utilizado 4 métricas de evaluación, dos métricas basadas en ejemplos como Hamming Loss y FMeasure ($FMeasure_{ex}$) y dos basadas en etiquetas, FMeasure micro ($FMeasure_{mic}$) y FMeasure macro ($FMeasure_{mac}$).

5.3.2. Resultados

En la Tabla 5.15 se recogen los rankings promedios obtenidos para cada método MLL y métrica considerando todas las parametrizaciones. De un primer análisis de los resultados, podemos observar que el algoritmo ECC obtuvo un mejor desempeño para 3 de las 4 métricas analizadas (Hamming Loss, $FMeasure_{ex}$ y $FMeasure_{mic}$), el algoritmo RAkEL obtuvo el mejor resultado para la métrica $FMeasure_{mac.}$ BR es el método que obtiene el tercer puesto para tres métricas (Hamming Loss, $FMeasure_{ex}$ y $FMeasure_{mic}$) y el segundo lugar para $FMeasure_{mac}$. El peor desempeño es para el algoritmo AdaBoost, seguido del algoritmo EPS. En resumen, según los resultados del Meta-Ranking, los tres métodos MLL que obtuvieron un mejor desempeño general fueron RAkEL, ECC y BR (ver Figura 5.9).

Métricas /Algoritmos	AdaBoost	BR	CC	ECC	EPS	HOMER	IBLR	LP	MlkNN	MLS	RAkEL
Hamming Loss	10.50	4.50	4.75	**1.25**	8.00	7.50	5.13	10.50	6.13	5.50	2.25
FMeasure$_{ex}$	11.00	3.63	4.25	**1.75**	10.00	5.00	7.00	9.00	8.00	4.38	2.00
Fmeasure$_{mic}$	11.00	3.75	3.88	**1.75**	10.00	5.38	7.00	9.00	8.00	4.38	1.88
Fmeasure$_{mac}$	11.00	3.13	3.25	4.25	10.00	4.13	8.00	7.00	9.00	4.38	**1.88**
Meta-Ranking	10.88	3.75	4.03	2.25	9.50	5.50	6.78	8.88	7.78	4.66	**2.00**

Tabla 5.15.- Resultados ranking promedio de los algoritmos multietiquetas analizados para cada métrica. En negrita se marcan los mejores resultados.

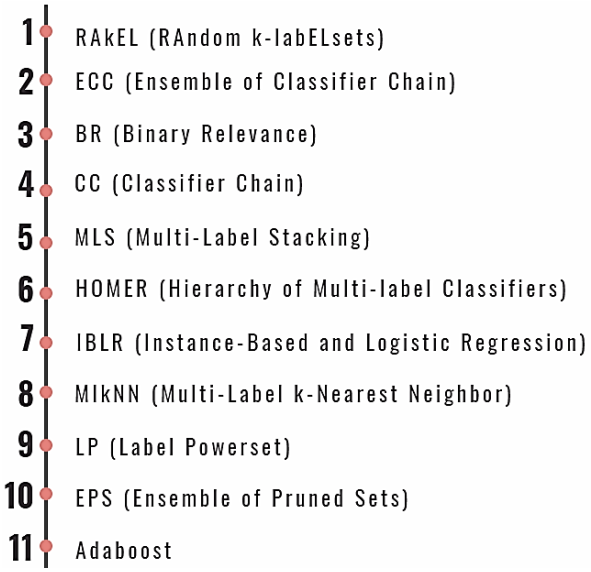


Figura 5.9- Ranking de métodos MLL en la tarea de clasificación diagnóstica (de mejor a peor desempeño)

En un segundo paso se aplicó un análisis estadístico basado en el test de ***Friedman*** [218] para poder rechazar la hipótesis nula, es decir, para poder rechazar la idea de que no existen diferencias entre los distintos métodos multietiqueta. Si el test de Friedman confirma que existen diferencias significativas, realizaremos en un post-hoc test, ***test de Shaffer*** [213], para poder identificar las diferencias existentes entre los distintos

algoritmos. En la Tabla 5.16 se recogen los resultados del test de Friedman para cada métrica. Como podemos comprobar, considerando una confianza del 95% (p-value menor de 0.05), se rechaza la hipótesis nula para todas las métricas de evaluación. Por tanto, una vez confirmado que se rechaza la hipótesis nula aplicaremos el ***test de Shaffer*** que nos permitirá realizar una comparativa múltiple entre los diferentes métodos multietiquetas (todos contra todos) para determinar si existen diferencias estadísticamente significativas entre ellos. En la Figura 5.10 se observan los resultados de dicho test (comparación múltiple de los distintos métodos multietiqueta para cada métrica).

Resultados test Friedman (métodos MLL)	
Métrica	p-value
Hamming loss	5.87E-10
FMeasure$_{ex}$	8.33E-12
Fmeasure$_{mic}$	5.49E-12
FMeasure$_{mac}$	3.41E-11

Tabla 5.16.- Test de Friedman para determinar si existen diferencias estadísticamente significativas entre los distintos métodos multietiqueta evaluados.

Centrándonos en la métrica Hamming Loss, el método que obtiene los mejores resultados es el ECC y tal como observamos en la Tabla 5.17 se aprecian diferencias significativas con los algoritmos Adaboost, EPS, HOMER y LP. Para la métrica *FMeasure$_{ex}$* el algoritmo que obtuvo mejores resultados fue también el algoritmo ECC, obteniendo diferencias significativas con los algoritmos Adaboost, EPS, LP y MLkNN (ver Tabla 5.18). Con respecto a la métrica *FMeasure$_{mic}$* el mejor desempeño lo arroja el algoritmo ECC, obteniendo diferencias significativas con los algoritmos AdaBoost, EPS, LP y MLkNN (ver Tabla 5.19). Por último, con respecto a la métrica *FMesaure$_{mac}$* observamos que el método con un mejor desempeño en la tarea de clasificación es para el algoritmo RakEL, el cual tiene diferencias significativas (con un nivel de confianza del 95%) con los algoritmos AdaBoost, EPS, IBLR y MLkNN (ver Tabla 5.20).

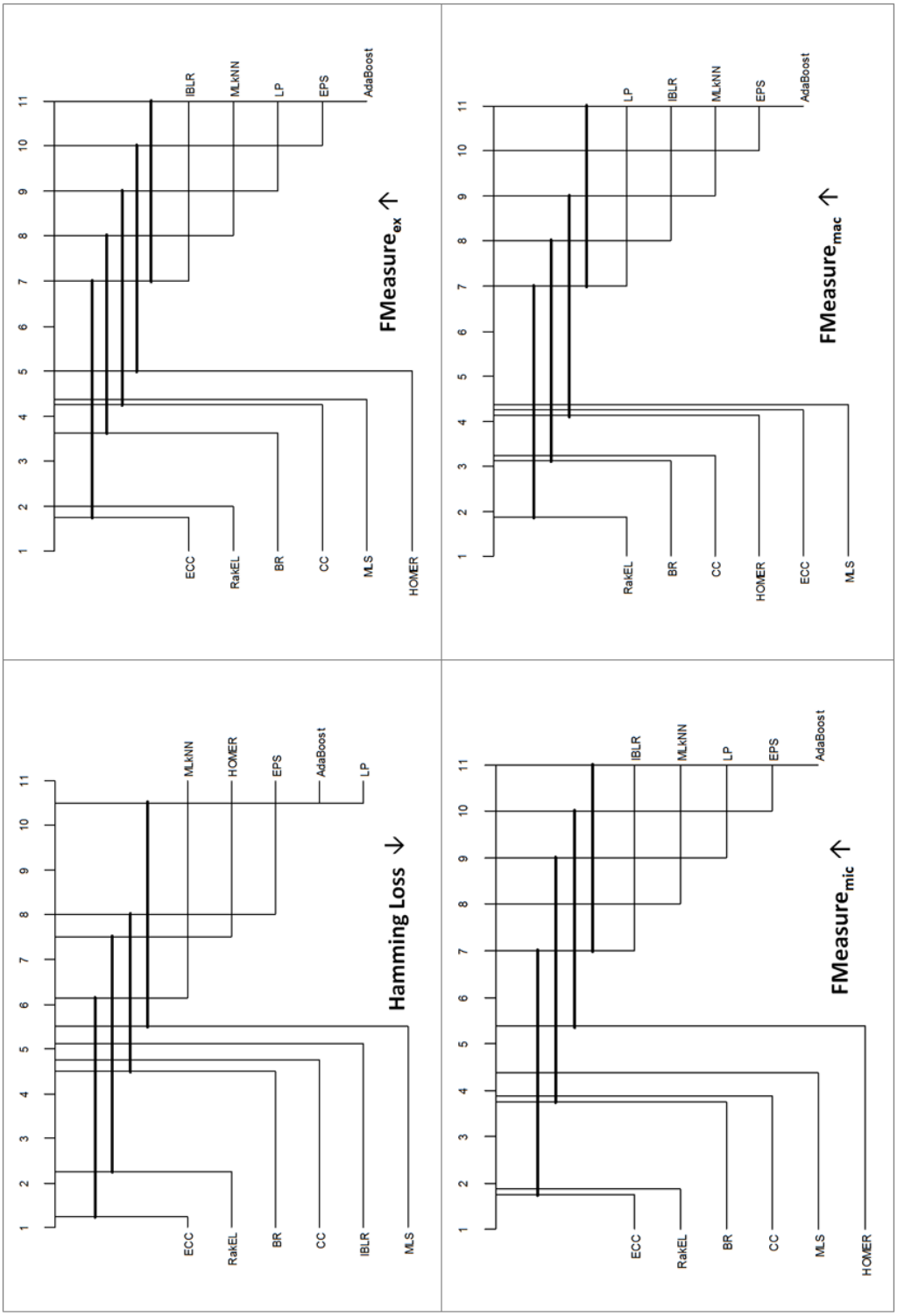


Figura 5.10. Comparación multiple de métodos multietiqueta para cada métrica usando el Test de Shaffer (nivel de confianza 95%).

Hamming Loss	AdaBoost	BR	CC	ECC	EPS	HOMER	IBLR	LP	MlkNN	MLS	RAkEL
AdaBoost	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	**TRUE**
BR	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE
CC	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE
ECC	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE
EPS	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**
HOMER	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
IBLR	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE
LP	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	**TRUE**
MLkNN	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
MLS	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
RakEL	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE

Tabla 5.17.- Comparativa múltiple de métodos multietiqueta mediante test de Shaffer para la métrica Hamming Loss. True: existen diferencias significativas/False: no existen diferencias significativas.

$FMeasure_{ex}$	AdaBoost	BR	CC	ECC	EPS	HOMER	IBLR	LP	MlkNN	MLS	RAkEL
AdaBoost	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
BR	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE
CC	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ECC	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	**TRUE**	**TRUE**	FALSE	FALSE
EPS	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
HOMER	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
IBLR	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
LP	FALSE	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**
MLkNN	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**
MLS	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
RakEL	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE

Tabla 5.18.- Comparativa múltiple de métodos multietiqueta mediante test de Shaffer para la métrica $FMeasure_{ex}$. True: existen diferencias significativas/False: no existen diferencias significativas.

Con respecto a la métrica $FMeasure_{mic}$ el mejor desempeño lo arroja el algoritmo ECC, obteniendo diferencias significativas con los algoritmos AdaBoost, EPS, LP y MLkNN (ver Tabla 5.19). Por último, con respecto a la métrica $FMesaure_{mac}$ observamos que el método con un mejor desempeño en la tarea de clasificación es para el algoritmo RakEL, el cual tiene diferencias significativas (con un nivel de confianza del 95%) con los algoritmos AdaBoost, EPS, IBLR y MLkNN (ver Tabla 5.20).

$FMeasure_{mic}$	**AdaBoost**	**BR**	**CC**	**ECC**	**EPS**	**HOMER**	**IBLR**	**LP**	**MlkNN**	**MLS**	**RAkEL**
AdaBoost	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
BR	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
CC	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ECC	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	**TRUE**	**TRUE**	FALSE	FALSE
EPS	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
HOMER	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
IBLR	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
LP	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**
MLkNN	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**
MLS	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
RakEL	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	**TRUE**	**TRUE**	FALSE	FALSE

Tabla 5.19.- Comparativa múltiple de métodos multietiqueta mediante test de Shaffer para la métrica $FMeasure_{mic}$. True: existen diferencias significativas/False: no existen diferencias significativas.

$FMeasure_{mac}$	**AdaBoost**	**BR**	**CC**	**ECC**	**EPS**	**HOMER**	**IBLR**	**LP**	**MlkNN**	**MLS**	**RAkEL**
AdaBoost	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
BR	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE
CC	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE
ECC	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
EPS	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	**TRUE**
HOMER	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
IBLR	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**
LP	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
MLkNN	FALSE	**TRUE**	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**
MLS	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
RakEL	**TRUE**	FALSE	FALSE	FALSE	**TRUE**	FALSE	**TRUE**	FALSE	**TRUE**	FALSE	FALSE

Tabla 5.20.- Comparativa múltiple de algoritmos multietiqueta mediante test de Shaffer para la métrica $FMeasure_{mac}$. True: existen diferencias significativas/False: no existen diferencias significativas.

5.3.3. Discusión

El tercer experimento se centró en determinar qué método MLL mejoraba el rendimiento de la tarea de clasificación diagnóstica multietiqueta. Para ello, se analizaron 11 métodos de clasificación multietiqueta del estado del arte como *AdaBoost, BR, CC, ECC, EPS, HOMER, IBLR, LP, MlkNN, MLS Y RAkEL.* De los resultados obtenidos en esta experimentación, y en base al test de Friedman, se confirmó que existían diferencias significativas entre los distintos métodos de clasificación analizados. En este análisis se verificó que los tres métodos que lograron un mejor rendimiento fueron RAkEL, ECC y BR. Sin embargo, no se encontraron diferencias estadísticamente significativas entre ellos (ver Figura 5.10) y por tanto, no pudimos afirmar estadísticamente cual era el mejor de los tres. Ante estos resultados, para el buen desempeño de la tarea de clasificación diagnóstica de MiNerDoc seleccionamos el método BR con SMO debido a una serie de ventajas de gran relevancia para la tarea propuesta: i) los buenos resultados obtenidos en la tarea de clasificación multietiqueta de acuerdo con el estado del arte, demostrando que el método BR es computacionalmente eficiente y efectivo, obteniendo un rendimiento competitivo frente a métodos más complejos cuando aumenta la complejidad del conjunto de datos (por ejemplo, mayor cardinalidad) [58]; ii) el rendimiento óptimo del método BR en escenarios clínicos específicos (por ejemplo, clasificación de enfermedades crónicas) [57]; iii) la simplicidad del método BR que, junto con el algoritmo SMO, garantiza buenos resultados en datasets que tienen un alto número de características y un bajo número de instancias (características que cumplen los datasets utilizados en nuestra experimentación) [229].

El método MLL seleccionado para llevar a cabo la tarea de clasificación diagnóstica incluida en el sistema MiNerDoc fue el método BR con SMO como clasificador base. Esta selección se ha basado en los óptimos resultados demostrados en la literatura relacionada con la clasificación multietiqueta junto con el análisis estadístico realizado en el que obtuvo el tercer puesto de once métodos evaluados (no existiendo diferencias significativas con los dos primeros métodos MLL).

5.4. Conclusiones generales

Es esta sección se resumen las principales conclusiones obtenidas de los distintos experimentos llevados a cabo para evaluar el sistema propuesto en esta tesis doctoral:

- El sistema MER de MiNerDoc, basado en diccionarios, obtuvo el primer puesto del ranking en las métricas FMeasure (81.54%) y Precisión (77.94%) frente a los sistemas CliNER y CLAMP.
- El Sistema MER de MiNerDoc aporta ventajas que lo diferencian de otros sistemas, permitiendo la detección de entidades negadas y la resolución de acrónimos.
- La metodología propuesta en esta investigación, dCSE (*diagnostic Classification with Semantic Enrichment*) para llevar a cabo la tarea CDA supera en rendimiento a la metodología Baseline.
- Existen diferencias significativas, con un 95% de confianza, entra las dos metodologías evaluadas (dCSE y Baseline) para las cuatro métricas analizadas.
- Las cuatro parametrizaciones construidas bajo la metodología dCSE han obtenido mejores resultados que las cuatro parametrizaciones construidas bajo la metodología Baseline (95% de confianza).
- La mejor combinación de técnicas de MT (parametrizaciones) para llevar a cabo la tarea CDA se ha dado con la parametrización dCSE-Param1. Esta parametrización se ha construido aplicando la técnica *stemmin*g (Porter), eliminando stop-words y realizando una tokenización basada en unigramas.
- La parametrización dCSE-Param1 ha obtenido diferencias estadísticamente significativas con respecto a todas las construidas bajo la metodología Baseline.
- La parametrización dCSE-Param1 se aplicada en la fase de ingeniería de características del sistema CDA de MiNerDoc.

- El uso de la tokenización basada en unigramas y la aplicación de la técnica *stemming* mejoran los resultados en la clasificación diagnóstica.
- Los tres mejores métodos MLL (de once métodos evaluados) fueron RakEL, ECC y BR, pero no se obtuvo diferencias estadísticamente significativas entre ellos.
- El modelo elegido que se aplica en el sistema CDA de MiNerDoc es el método BR junto con el clasificador de base SMO. Esta elección se basó en los buenos resultados obtenidos en el análisis experimental y en los buenos resultados descritos en la literatura, que consideran esta combinación óptima y eficaz para la tarea propuesta.
- El sistema CDA de MiNerDoc (siguiendo la metodología dCSE, la parametrización dCSE-Param1 y el método BR) puede alcanzar valores del 79.88% para la métrica $FMeasure_{ex}$, 81.04% para $FMeasure_{mic}$ y un 66.47% para la métrica $FMeasure_{mac}$.
- En resumen, se ha podido verificar, en base a los distintos experimentos realizados, el prometedor desempeño de MiNerDoc en las dos tareas evaluadas, reconocimiento de entidades médicas (FMeasure 81.54%) y clasificación diagnóstica ($FMeasure_{mic}$ 81,04%).

6

Conclusiones y líneas de trabajos futuros

"La conclusión es que sabemos muy poco y sin embargo es asombroso lo mucho que conocemos. Y más asombroso todavía que un conocimiento tan pequeño pueda dar tanto poder."

Bertrand Arthur William Rusell

En este Capítulo presentamos un resumen de los objetivos alcanzados durante el desarrollo de esta tesis doctoral, detallando las principales conclusiones y resultados obtenidos. Posteriormente se recogen las nuevas líneas de investigación que pueden derivarse del presente trabajo.

6.1. Conclusiones finales

El principal objetivo de esta tesis doctoral se centraba en desarrollar y poner en práctica una metodología, basada en el área de la Minería de Textos (MT), capaz de transformar la información clínica textual en conocimiento con el fin de apoyar al profesional sanitario en la toma de decisiones y la detección temprana. El desarrollo de este objetivo ha dado como resultado la creación de un sistema, denominado MiNerDoc, que permite resolver automáticamente tareas complejas (detección de factores de riesgo y predicción diagnóstica) para facilitar la prevención primaria y la toma decisiones clínicas. A continuación, resumiremos los resultados y logros más destacados derivados de este trabajo.

- **Revisión y análisis bibliográfico.** Se ha realizado un extenso estudio y revisión bibliográfica relacionada con los temas en los que se centra esta tesis, como MT, AA, PLN, clasificación diagnóstica, reconocimiento de entidades, MetaMap, UMLS, etc. Se han analizado y recopilado más de 270 artículos, la mayoría de estas publicaciones se centran en el ámbito de la Medicina.

- **Creación de una metodología para el Reconocimiento de Entidades Médicas.** Se ha desarrollado una metodología, basada en técnicas de MT, técnicas del PLN y la herramienta MetaMap, para construir un sistema de reconocimiento y extracción de entidades médicas. Este modelo es capaz de detectar y extraer cinco tipos diferentes de entidades médicas (*disease, region/part body, pharmacologic, procedure/test, finding/sign*). En base a los resultados obtenidos en la fase experimental, hemos demostrado que la metodología propuesta, basada en diccionarios, puede ser competitiva con otros modelos basados en AA o en reglas.

- **Creación de una nueva metodología (denominada dCSE) aplicada a la clasificación diagnóstica multietiqueta.** Hemos incorporado fuentes externas de conocimiento como la herramienta MetaMap y UMLS, para mejorar el rendimiento de la clasificación diagnóstica. Estos recursos nos han permitido resolver ambigüedades terminológicas, expandir el vocabulario con sinónimos, identificar acrónimos, detectar negaciones, entre otras funcionalidades. En base a distintos test estadísticos hemos demostrado que esta metodología mejora el rendimiento de la clasificación diagnóstica frente a los métodos convencionales que no aplican enriquecimiento semántico.

- **Creación de un novedoso sistema de MT, denominado MiNerDoc.** Hemos comprobado, a lo largo del desarrollo de esta tesis, que existen muy pocos

sistemas que den valor al texto para facilitar el trabajo clínico mediante la automatización de tareas complejas. En este sentido, se propuso el desarrollo de un nuevo sistema de MT, denominado MiNerDoc, cuyo objetivo principal era apoyar el proceso de toma de decisiones clínicas mediante el análisis de informes clínicos textuales bajo un *framework* unificado que realiza múltiples funcionalidades. Su principal objetivo es facilitar tareas complejas como la detección de factores de riesgo y la predicción automática de códigos de diagnósticos normalizados. MiNerDoc integra una combinación de técnicas de la MT junto con el enriquecimiento terminológico y semántico proporcionado por la herramienta MetaMap y el metatesauro UMLS. Un resumen de las principales funcionalidades del sistema MiNerDoc, que han sido ampliamente detalladas a lo largo del presente trabajo, son: i) Detección de cinco tipos de entidades médicas (*Disease, Pharmacologic, Region/Part Body, Procedure/Test, Finding/Sign*); ii) Detección de factores de riesgo, iii) Detección de Negaciones; iv) Predicción automática de códigos normalizados de diagnóstico (de uno o múltiples informes clínicos).

- **Evaluación del sistema MiNerdoc**. Hemos evaluado MiNerDoc acercando su funcionamiento a un entorno similar a un escenario real. Para ello, hemos utilizado un conjunto de informes de alta de pacientes reales tomados de la colección MIMIC (datos con deidentificación). Los resultados demostraron la efectividad y viabilidad del sistema propuesto y verificaron el prometedor rendimiento de MiNerDoc en las dos tareas evaluadas, reconocimiento de entidades médicas (FMeasure 81.54%) y clasificación diagnóstica multietiqueta ($FMeasure_{mic}$ 81.04%).

6.2. Líneas de trabajo futuras

A lo largo del desarrollo de esta tesis doctoral, hemos creado un sistema, basado en MT y MetaMap, capaz de abordar tareas como el reconocimiento de entidades médicas y la clasificación diagnóstica. Llegados a este punto, hemos observado que todavía queda un largo camino por recorrer y que son muchas las líneas de investigación que podrían derivarse de las ya desarrolladas. Algunas de estas líneas futuras se recogen a continuación:

- La adaptación de las metodologías desarrolladas en esta investigación para aplicarlas a documentos clínicos escritos en castellano, sería una de las nuevas líneas de investigación que sería necesaria abrir. En este caso, deberían incorporarse otras herramientas y fuentes de conocimiento externo como FreeLing, GATE y SNOMED-CT.

- Con respecto al sistema MER, sería necesario abrir una nueva vía para investigar el desarrollo de una solución híbrida basada en diccionarios y aprendizaje automático, donde el problema de la identificación de la entidad se convierta en un problema de clasificación pero con el enriquecimiento que pueda aportar la utilización de diccionarios específicos del dominio.

- El sistema de clasificación diagnóstica multietiqueta desarrollado en esta tesis se ha basado en la categorización de informes clínicos según códigos normalizados de diagnóstico MeSH. Una futura línea de investigación, de gran interés en los entornos hospitalarios y en especial para las unidades de documentación clínica, podría ser la realización de un sistema que permita la predicción automática de códigos de diagnóstico según la Clasificación Internacional de Enfermedades (CIE), en concreto la CIE-10, al tratarse de una de las clasificaciones más complejas y extensas.

- La investigación sobre clasificación diagnóstica multietiqueta se ha realizado sobre técnicas de representación de textos basada en el modelo de bolsa de palabras. Sería interesante abrir una nueva vía de investigación incorporando nuevos enfoques como el basado en el modelo Word2Vec o Doc2Vec.

- La aplicación MiNerDoc lleva a cabo dos tareas básicas, la detección de entidades médicas y la clasificación multietiqueta. Una tarea importante en el ámbito médico, y que abriría una nueva línea de investigación, sería la incorporación de la detección automática de relaciones entre entidades médicas (interrelaciones entre tratamientos y síntomas, interacciones farmacológicas, relaciones entre diagnóstico y síntomas, etc).

- Debido al gran volumen de documentación clínica, en forma de registros electrónicos, que se genera en los centros sanitarios sería necesaria iniciar una línea de investigación centrada en combinar la tecnología de MT junto con la tecnología *Big Data*, con el objetivo de llevar a cabo el procesamiento masivo de grandes colecciones de informes clínicos.

7

PUBLICACIONES ASOCIADAS A LA TESIS

7.1. Revistas internacionales

1. TÍTULO: An advanced review on text mining in medicine.

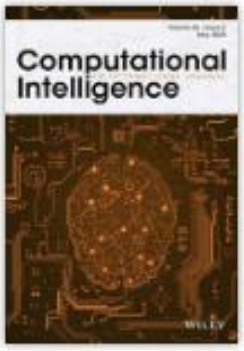
AUTORES: Carmen Luque, José M. Luna, María Luque, Sebastián Ventura

Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, vol. 9, no 3, p. e1302.

DOI: 10.1002/widm.1302.

2. **TÍTULO: A semantically enriched text mining system for clinical decision support**

AUTORES: Carmen Luque, José M. Luna, Sebastián Ventura



Computational Intelligence. 2020; 1–26.

DOI: 10.1111/coin.12322

7.2. Conferencias internacionales

1. **TÍTULO. MiNerDoc: a Semantically Enriched Text Mining System to Transform Clinical Text into Knowledge**

AUTORES. Carmen Luque, José M. Luna, Sebastián Ventura

IEEE Xplore® Digital Library

2019 IEEE 32nd International Symposium on Computer-Based Medical Systems

CBMS 2019

IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2019. p. 702-707.

DOI: 10.1109/CBMS.2019.00142

BIBLIOGRAFÍA

[1] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

[2] Moreno, L., Palomar, M., Molina, A., & Ferrández, A. (1999). Introducción al procesamiento del lenguaje natural. *Servicio de Publicaciones Universidad de Alicante. Universidad de Alicante*.

[3] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261-266.

[4] Hernández, M. B., & Gómez, J. M. (2013). Aplicaciones de procesamiento de lenguaje natural. *Revista Politécnica*, *32*.

[5] Shaalan, K., Hassanien, A. E., & Tolba, F. (Eds.). (2017). *Intelligent Natural Language Processing: Trends and Applications* (Vol. 740). Springer.

[6] Rojas, Y., Ferrández, A., & Peral, J. (2005). Aplicación del Procesamiento de Lenguaje Natural en la Recuperación de Información. *Procesamiento del lenguaje natural*, *34*.

[7] Shah, U. S., & Jinwala, D. C. (2015). Resolving ambiguities in natural language software requirements: a comprehensive survey. *ACM SIGSOFT Software Engineering Notes*, *40*(5), 1-7.

[8] Jiménez Zafra, S. M., Martínez Cámara, E., Martín Valdivia, M. T., & Molina González, M. D. (2015). Tratamiento de la Negación en el Análisis de Opiniones en Espanol. *Procesamiento del Lenguaje Natural*, (54).

[9] Dalianis, H. (2018). Basic Building Blocks for Clinical Text Processing. In *Clinical Text Mining* (pp. 55-82). Springer, Cham.

[10] Marrero, M., Sánchez-Cuadrado, S., Urbano, J., Morato, J., & Moreiro, J. A. (2009). Sistemas de recuperación de información adaptados al dominio biomédico. *El profesional de la información*, *19*(3), 246-254.

[11] Dale, R., Moisl, H., & Somers, H. (2000). *Handbook of natural language processing*. CRC Press.

[12] Wu, S. T., Kaggal, V. C., Dligach, D., Masanz, J. J., Chen, P., Becker, L., ... & Chute, C. G. (2013). A common type system for clinical natural language processing. *Journal of biomedical semantics*, *4*(1), 1.

[13] Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. (2016). Natural language processing in radiology: a systematic review. *Radiology*, *279*(2), 329-343.

[14] Baeza-Yates, R., & Ribeiro, B. D. A. N. (2011). *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley,

[15] Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, *39*(1), 80-91.

[16] Collier, N., Park, H. S., Ogata, N., Tateishi, Y., Nobata, C., Ohta, T., ... & Tsujii, J. I. (1999, June). The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* (pp. 271-272). Association for Computational Linguistics.

[17] Vilares, J. (2006). Aplicaciones del Procesamiento del Lenguaje Natural en la Recuperación de Información en Español. *Procesamiento del lenguaje natural*, *36*.

[18] Hirschman, L. (1998). The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech & Language*, *12*(4), 281-305.

[19] Merchant, R. H. (1993). Tipster program overview. In *TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, Virginia, September 19-23, 1993*.

[20] Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., & Tyson, M. (1997). FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. *Finite-state language processing*, 383-406.

[21] Soderland, S., Fisher, D., Aseltine, J., & Lehnert, W. (1995). CRYSTAL: Inducing a conceptual dictionary. *arXiv preprint cmp-lg/9505020*.

[22] Riloff, E. (1993, July). Automatically constructing a dictionary for information extraction tasks. In *AAAI* (Vol. 1, No. 1, pp. 2-1).

[23] Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval* (Vol. 1). Cambridge: MIT press.

[24] Gutiérrez-Artacho, J., & Olvera Lobo, M. D. (2017). An Overview of the Linguistic Resources used in Cross-Language Question Answering Systems in CLEF Conference.

[25] Piskorski, J., & Yangarber, R. (2013). Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization* (pp. 23-49). Springer, Berlin, Heidelberg.

[26] Cardie, C. (1997). Empirical methods in information extraction. *AI magazine*, *18*(4), 65.

[27] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

[28] Dietterich, T. G. (1997). Machine-learning research. *AI magazine*, *18*(4), 97.

[29] Mitchell, T. M. (2006). *The discipline of machine learning* (Vol. 9). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[30] Simon, H. A. (1983). Why should machines learn?. In *Machine Learning, Volume I* (pp. 25-37).

[31]Langley, P. (2011). The changing science of machine learning. *Machine Learning*, *82*(3), 275-279.

[32] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89-109.

[33] Wang, S., & Summers, R. M. (2012). Machine learning and radiology. *Medical image analysis*, *16*(5), 933-951.

[34] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[35] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

[36] Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321.

[37] Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, *37*(2), 505-515.

[38] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

[39]Gilchrist, A. (2003). Thesauri, taxonomies and ontologies–an etymological note. *Journal of documentation*, *59*(1), 7-18.

[40] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12).

[41] Bassiouni, M., Ali, M., & El-Dahshan, E. A. (2018). Ham and Spam E-Mails Classification Using Machine Learning Techniques. *Journal of Applied Security Research*, *13*(3), 315-331.

[42] Malhotra, A., Younesi, E., Gündel, M., Müller, B., Heneka, M. T., & Hofmann-Apitius, M. (2014). ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's & dementia*, *10*(2), 238-246.

[43] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1* (pp. 173-180). Association for computational Linguistics.

[44] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[45] Celebi, M. E., & Aydin, K. (Eds.). (2016). *Unsupervised Learning Algorithms*. Springer.

[46] Anitha, P., & Patil, M. M. (2019). RFM model for Customer Purchase Behavior using K-Means Algorithm. *Journal of King Saud University-Computer and Information Sciences*.

[47] Pondel, M., & Korczak, J. (2018, September). Collective Clustering of Marketing Data-Recommendation System Upsaily. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 801-810). IEEE.

[48] Shash, S. F., & Mollá, D. (2013, May). Clustering of medical publications for evidence based medicine summarisation. In *Conference on Artificial Intelligence in Medicine in Europe* (pp. 305-309). Springer, Berlin, Heidelberg.

[49] Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014, February). DBSCAN: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)* (pp. 232-238). IEEE.

[50] Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Frontiers in computational neuroscience*, *13*.

[51] De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In *Lrec* (Vol. 6, pp. 449-454).

[52] He, Y., Sarntivijai, S., Lin, Y., Xiang, Z., Guo, A., Zhang, S., ... & Smith, B. (2014). OAE: the ontology of adverse events. Journal of biomedical semantics, 5(1), 29.

[53] Miyao, Y., & Tsujii, J. I. (2005, June). Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 83-90). Association for Computational Linguistics.

[54] Rector, A., & Rogers, J. (2004). Patterns, properties and minimizing commitment: Reconstruction of the galen upper ontology in owl. In *Proceedings of the EKAW* (Vol. 4).

[55] Moody, G. B., & Mark, R. G. (1996, September). A database to support development and evaluation of intelligent intensive care monitoring. In *Computers in Cardiology 1996* (pp. 657-660). IEEE.

[56] Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

[57] Zufferey, D., Hofer, T., Hennebert, J., Schumacher, M., Ingold, R., & Bromuri, S. (2015). Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Computers in biology and medicine*, *65*, 34-43.

[58] Luaces, O., Díez, J., Barranquero, J., del Coz, J. J., & Bahamonde, A. (2012). Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, *1*(4), 303-313.

[59] Roberts, K., Rink, B., & Harabagiu, S. M. (2013). A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association*, *20*(5), 867-875.

[60] McCallum, A. K. (2002). Mallet: A machine learning for language toolkit (2002).

[61] Chapelle, O., Scholkopf, B., & Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, *20*(3), 542-542.

[62] Balcan, M. F., Blum, A., Choi, P. P., Lafferty, J., Pantano, B., Rwebangira, M. R., & Zhu, X. (2005, August). Person identification in webcam images: An application of semi-supervised learning. In *ICML 2005 Workshop on Learning with Partially Classified Training Data* (Vol. 2, p. 6).

[63] Liu, J., Chen, C., Bu, J., You, M., & Tao, J. (2007, July). Speech emotion recognition using an enhanced co-training algorithm. In *Multimedia and Expo, 2007 IEEE International Conference on*(pp. 999-1002). IEEE.

[64] Frinken, V., Fischer, A., Bunke, H., & Foornes, A. (2011, September). Co-training for handwritten word recognition. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (pp. 314-318). IEEE.

[65] Màrquez, L., & Rodríguez, H. (1998, April). Part-of-speech tagging using decision trees. In *European Conference on Machine Learning* (pp. 25-36). Springer, Berlin, Heidelberg.

[66] Jovic, A., Prcela, M., & Gamberger, D. (2007, June). Ontologies in medical knowledge representation. In *2007 29th International Conference on Information Technology Interfaces* (pp. 535-540). IEEE.

[67] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, *33*(suppl_1), D514-D517.

[68] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[69] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.

[70] Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space?. *Machine Learning*, *46*(1-3), 423-444.

[71] Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J. A., & Plaza, A. (2015). On understanding big data impacts in remotely sensed image classification using support

vector machine methods. *IEEE journal of selected topics in applied earth observations and remote sensing*, *8*(10), 4634-4646.

[72] Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W. W., Mazaitis, M., Felix, V., ... & Kibbe, W. A. (2011). Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, *40*(D1), D940-D946.

[73]Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

[74] Leaman, R., Miller, C., & Gonzalez, G. (2009, November). Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. In *Proceedings of the 2009 Symposium on Languages in Biology and Medicine* (Vol. 82, No. 9).

[75]Carreras, X., Chao, I., Padró, L., & Padró, M. (2004, May). FreeLing: An Open-Source Suite of Language Analyzers. In *LREC* (pp. 239-242).

[76] Màrquez, L., & Giménez, J. (2004). A general pos tagger generator based on support vector machines. *Journal of Machine Learning Research*.

[77]Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., & Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, *8*(1), 50.

[78] Rodríguez, C. I. L., Sánchez, M. T., & Benítez, P. F. (2006). Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm. *RevistaeSalud. com*, *2*(8).

[79] Sucar, L. E., & Tonantzintla, M. (2006). Redes bayesianas. *BS Araujo, Aprendizaje Automático: conceptos básicos y avanzados*, 77-100.

[80]Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*(3), 5432-5435.

[81] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000, July). An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 160-167). ACM.

[82] Chen, H., & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC bioinformatics*, *5*(1), 147.

[83] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

[84] Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

[85] Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3-26.

[86] Hettne, K. M., Stierum, R. H., Schuemie, M. J., Hendriksen, P. J., Schijvenaars, B. J., Mulligen, E. M. V., ... & Kors, J. A. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, *25*(22), 2983-2991.

[87] Garla, V. N., & Brandt, C. (2012). Ontology-guided feature engineering for clinical text classification. *Journal of biomedical informatics*, *45*(5), 992-998.

[88] Waraporn, P., Meesad, P., & Clayton, G. (2010). Ontology-supported processing of clinical text using medical knowledge integration for multi-label classification of diagnosis coding. *arXiv preprint arXiv:1004.1230*.

[89] Zhang, K., Ma, H., Zhao, Y., Zan, H., & Zhuang, L. (2018). The Comparative Experimental Study of Multilabel Classification for Diagnosis Assistant Based on Chinese Obstetric EMRs. *Journal of healthcare engineering*, *2018*.

[90] Erraguntla, M., Gopal, B., Ramachandran, S., & Mayer, R. (2012, January). Inference of missing ICD 9 codes using text mining and nearest neighbor techniques. In *2012 45th Hawaii International Conference on System Sciences* (pp. 1060-1069). IEEE.

[91] McCulloch, W. S., & Pitts, W. (1990). A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biology*, *52*(1-2), 99-115.

[92] Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, *47*(11), 1940-1951.

[93] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

[94] Song, M., Yu, H., & Han, W. S. (2015). Developing a hybrid dictionary-based bio-entity recognition technique. *BMC medical informatics and decision making*, *15*(1), S9.

[95] Tsuruoka, Y., & Tsujii, J. I. (2003, July). Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13* (pp. 41-48). Association for Computational Linguistics.

[96] Kavuluru, R., Rios, A., & Lu, Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, *65*(2), 155-166.

[97] Dande, P., & Samant, P. (2018). Acquaintance to artificial neural networks and use of artificial intelligence as a diagnostic tool for tuberculosis: a review. *Tuberculosis*, *108*, 1-9.

[98] Peng, J., Estrada, G., Pedersoli, M., & Desrosiers, C. (2020). Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107269.

[99] Tsuruoka, Y., Tsujii, J. I., & Ananiadou, S. (2008). FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, *24*(21), 2559-2560.

[100] Leser, U., & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in bioinformatics*, *6*(4), 357-369.

[101] Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, *8*(2), 339-344.

[102] Shaalan,K.(2010).Rule-based approach in Arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, *3*(3), 11-19.

[103] Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of predictive text mining*. Springer.

[104] Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A brief survey of text mining. In *Ldv Forum* (Vol. 20, No. 1, pp. 19-62).

[105] Verma, T., Renu, R., & Gaur, D. (2014). Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems*, *7*(2), 16-18.

[106] Silva, C., & Ribeiro, B. (2003, July). The importance of stop word removal on recall values in text categorization. In *Neural Networks, 2003. Proceedings of the International Joint Conference on* (Vol. 3, pp. 1661-1666). IEEE.

[107] Lourdusamy, R., & Abraham, S. (2018). A Survey on Text Pre-processing Techniques and Tools. *International Journal of Computer Sciences and Engineering Open Access Survey Paper*, *6*.

[108] Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004, November). Stemming and lemmatization in the clustering of finnish text documents. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 625-633). ACM.

[109] Rani, S. R., Ramesh, B., Anusha, M., & Sathiaseelan, J. G. R. (2015). Evaluation of stemming techniques for text classification. *International Journal of Computer Science and Mobile Computing*, *4*(3), 165-171.

[110] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137.

[111]Leaman, R., Islamaj Doğan, R., & Lu, Z. (2013). DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, *29*(22), 2909-2917.

[112] Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances.

[113] Ferraro, J. P., Daumé III, H., DuVall, S. L., Chapman, W. W., Harkema, H., & Haug, P. J. (2013). Improving performance of natural language processing part-of-speech

tagging on clinical narratives through domain adaptation. *Journal of the American Medical* tagging on *Informatics Association*, *20*(5), 931-939.

[114] Cotik, V., Roller, R., Xu, F., Uszkoreit, H., Budde, K., & Schmidt, D. (2016). Negation detection in clinical reports written in German. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*(pp. 115-124).

[115] Mehrabi, S., Krishnan, A., Sohn, S., Roch, A. M., Schmidt, H., Kesterson, J., ... & Palakal, M. (2015). DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of biomedical informatics*, *54*, 213-219.

[116] Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

[117] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

[118] Wang, F., Li, C. H., Wang, J. S., Xu, J., & Li, L. (2015). A two-stage feature selection method for text categorization by using category correlation degree and latent semantic indexing. *Journal of Shanghai Jiaotong University (Science)*, *20*(1), 44-50.

[119] Ju, R., Zhou, P., Li, C. H., & Liu, L. (2015, October). An efficient method for document categorization based on word2vec and latent semantic analysis. In *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on* (pp. 2276-2283). IEEE.

[120] Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52-61.

[121] Borisov, A., Serdyukov, P., & de Rijke, M. (2016, April). Using Metafeatures to Increase the Effectiveness of Latent Semantic Models in Web Search. In*Proceedings of the 25th International Conference on World Wide Web* (pp. 1081-1091). International World Wide Web Conferences Steering Committee.

[122] Kireyev, K. (2008). Using Latent Semantic Analysis for Extractive Summarization. In *TAC*.

[123] Steinberger, J., & Křišťan, M. (2007). Lsa-based multi-document summarization. In *Proceedings of 8th International PhD Workshop on Systems and Control* (Vol. 7).

[124] Domeniconi, G., Moro, G., Pasolini, R., & Sartori, C. (2015, July). A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf. idf. In *DATA* (pp. 26-37).

[125] Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38*(7), 9014-9022.

[126] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

[127] Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1-4), 43-52.

[128] Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, *31*(4), 721-735.

[129] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

[130] Oren, N. (2002, September). Reexamining tf. idf based information retrieval with genetic programming. In *Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology* (pp. 224-234). South African Institute for Computer Scientists and Information Technologists.

[131] Sahlgren, M., & Cöster, R. (2004, August). Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 487). Association for Computational Linguistics.

[132] Forman, G. (2008, October). BNS feature scaling: an improved representation over tf-idf for svm text classification. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 263-270). ACM.

[133] Roul, R. K., Devanand, O. R., & Sahay, S. K. (2014). Web document clustering and ranking using tf-idf based apriori approach. *arXiv preprint arXiv:1406.5617*.

[134] Gelgi, F., Davulcu, H., & Vadrevu, S. (2007, June). Term Ranking for Clustering Web Search Results. In *WebDB*.

[135] Carrera-Trejo, J. V., Sidorov, G., Miranda-Jiménez, S., Moreno Ibarra, M., & Cadena Martínez, R. (2015). Latent Dirichlet Allocation complement in the vector space model for Multi-Label Text Classification. *International Journal of Combinatorial Optimization Problems and Informatics*, *6*(1), 7-19.

[136] Monteiro, L. B., Weigang, L., & Saleh, A. A. (2015, October). An Approach of Vector Space Model to Link Concrete Concepts with Wiki Entities. In *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on* (pp. 313-320). IEEE.

[137] Chen, H., Fuller, S. S., Friedman, C., & Hersh, W. (2005). Knowledge management, data mining, and text mining in medical informatics. In *Medical Informatics* (pp. 3-33). Springer, Boston, MA.

[138] Huang, C. C., & Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, *17*(1), 132-144.

[139] Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, *88*(3), 265.

[140] Wang, Y., & Patrick, J. (2009, September). Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the workshop on biomedical information extraction* (pp. 42-49). Association for Computational Linguistics.

[141] Valenzuela-Escárcega, M. A., Hahn-Powell, G., Surdeanu, M., & Hicks, T. (2015). A domain-independent rule-based framework for event extraction. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, 127-132.

[142] Abacha, A. B., & Zweigenbaum, P. (2011, June). Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop* (pp. 56-64). Association for Computational Linguistics.

[143] Roberts, K., & Harabagiu, S. M. (2011). A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, *18*(5), 568-573.

[144] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[145] Aparicio, F., de Buenaga, M., Rubio, M., Hernando, M. A., Gachet, D., Puertas, E., & Giráldez, I. (2010). TMT: A tool to guide users in finding information on clinical texts. *Procesamiento del lenguaje natural*, *46*, 27-34.

[146] Kipper-Schuler, K., Kaggal, V., Masanz, J., Ogren, P., & Savova, G. (2008). System evaluation on a named entity corpus from clinical notes. In *Language resources and evaluation conference, LREC* (pp. 3001-3007).

[147] Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., & Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, *18*(5), 601-606.

[148] Xu, Y., Tsujii, J., & Chang, E. I. C. (2012). Named entity recognition of follow-up and time information in 20 000 radiology reports. *Journal of the American Medical Informatics Association*, *19*(5), 792-799.

[149] Gao, J., Liu, N., Lawley, M., & Hu, X. (2017). An interpretable classification framework for information extraction from online healthcare forums. *Journal of healthcare engineering*, *2017*.

[150] Bodnari, A., Deleger, L., Lavergne, T., Neveol, A., & Zweigenbaum, P. (2013, September). A Supervised Named-Entity Extraction System for Medical Text. In *CLEF (Working Notes)*.

[151] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507-513.

[152] Xia, Y., Zhong, X., Liu, P., Tan, C., Na, S., Hu, Q., & Huang, Y. (2013, September). Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1. In *CLEF (Working Notes)*.

[153] Skeppstedt, M., Kvist, M., Nilsson, G. H., & Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, *49*, 148-158.

[154] Carrero, F., Cortizo, J. C., & Gómez, J. M. (2008, November). Building a Spanish MMTx by using automatic translation and biomedical ontologies. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 346-353). Springer, Berlin, Heidelberg.

[155] Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229-236.

[156] Lin, Y. K., Chen, H., & Brown, R. A. (2013). MedTime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics*, *46*, S20-S28.

[157] Ferrández, O., South, B. R., Shen, S., & Meystre, S. M. (2012, June). A hybrid stepwise approach for de-identifying person names in clinical documents. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing* (pp. 65-72). Association for Computational Linguistics.

[158] Benton, A., Hill, S., Ungar, L., Chung, A., Leonard, C., Freeman, C., & Holmes, J. H. (2011). A system for de-identifying medical message board text. *BMC bioinformatics*, *12*(3), S2.

[159] Krupka, G. R., & Hausman, K. (1998, April). Isoquest inc.: Description of the netowl (tm) extractor system as used for muc-7. In *Proceedings of MUC* (Vol. 7).

[160] Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, *6*(1), 57-71.

[161] Lee, K., Kim, B., Choi, Y., Kim, S., Shin, W., Lee, S., ... & Kang, J. (2018). Deep learning of mutation-gene-drug relations from the literature. *BMC bioinformatics*, *19*(1), 21.

[162] Van Landeghem, S., Björne, J., Abeel, T., De Baets, B., Salakoski, T., & Van de Peer, Y. (2012, June). Semantically linking molecular entities in literature through entity relationships. In *BMC bioinformatics* (Vol. 13, No. 11, p. S6). BioMed Central.

[163] Zhou, J., & Fu, B. Q. (2018). The research on gene-disease association based on text-mining of PubMed. *BMC bioinformatics*, *19*(1), 37.

[164] Galustian, C., & Dalgleish, A. G. (2010). The power of the web in cancer drug discovery and clinical trial design: research without a laboratory?. *Cancer informatics*, *9*, 31-5.

[165] Selvaraj, S., & Natarajan, J. (2011). Microarray data analysis and mining tools. *Bioinformation*, *6*(3), 95.

[166] Uzuner, O., Mailoa, J., Ryan, R., & Sibanda, T. (2010). Semantic relations for problem-oriented medical records. *Artificial intelligence in medicine*, *50*(2), 63-73.

[167] Chang, Y. C., Dai, H. J., Wu, J. C. Y., Chen, J. M., Tsai, R. T. H., & Hsu, W. L. (2013). TEMPTING system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. *Journal of biomedical informatics*, *46*, S54-S62.

[168] Zhu, X., Cherry, C., Kiritchenko, S., Martin, J., & De Bruijn, B. (2013). Detecting concept relations in clinical text: Insights from a state-of-the-art model. *Journal of biomedical informatics*, *46*(2), 275-285.

[169] Rink, B., Harabagiu, S., & Roberts, K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, *18*(5), 594-600.

[170] Wright, A., Chen, E. S., & Maloney, F. L. (2010). An automated technique for identifying associations between medications, laboratory results and problems. *Journal of biomedical informatics*, *43*(6), 891-901.

[171] Munkhdalai, T., Liu, F., & Yu, H. (2018). Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. *JMIR public health and surveillance*, *4*(2), e29.

[172] Ningthoujam, D., Yadav, S., Bhattacharyya, P., & Ekbal, A. (2019). Relation extraction between the clinical entities based on the shortest dependency path based LSTM. *arXiv preprint arXiv:1903.09941*.

[173] Chikka, V. R., & Karlapalem, K. (2018). A hybrid deep learning approach for medical relation extraction. *arXiv preprint arXiv:1806.11189*.

[174] Black, W. J., Rinaldi, F., & Mowatt, D. (1998). FACILE: Description of the NE System Used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

[175] Swanson, D. R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science*, *38*(4), 228-233.

[176] Smalheiser, N. R., & Swanson, D. R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *Neuroscience research communications*, *15*(1), 1-9.

[177] Byrd, R. J., Steinhubl, S. R., Sun, J., Ebadollahi, S., & Stewart, W. F. (2014). Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International journal of medical informatics*, *83*(12), 983-992.

[178] Perez, A., Weegar, R., Casillas, A., Gojenola, K., Oronoz, M., & Dalianis, H. (2017). Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. Journal of biomedical informatics, 71, 16-30.

[179] Collier, N. (2012). Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global public health*, *7*(7), 731-749.

[180] Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., ... & Shigematsu, M. (2008). BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, *24*(24), 2940-2941.

[181] Baron, J. A., Senn, S., Voelker, M., Lanas, A., Laurora, I., Thielemann, W.,& McCarthy, D. (2013). Gastrointestinal adverse effects of short-term aspirin use: a meta-analysis of published randomized controlled trials. *Drugs in R&D*, *13*(1), 9-16.

[182] Tafti, A. P., Badger, J., LaRose, E., Shirzadi, E., Mahnke, A., Mayer, J., ... & Peissig, P. (2017). Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR medical informatics*, *5*(4).

[183] Yang, Y., Xie, P., Gao, X., Cheng, C., Li, C., Zhang, H., & Xing, E. (2017). Predicting Discharge Medications At Admission Time Based On Deep Learning. *arXiv preprint arXiv:1711.01386*.

[184] Heintzelman, N. H., Taylor, R. J., Simonsen, L., Lustig, R., Anderko, D., Haythornthwaite, J. A., ... & Bova, G. S. (2012). Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *Journal of the American Medical Informatics Association*, *20*(5), 898-905.

[185] Bhat, A., Shih, G., & Zabih, R. (2011, August). Automatic selection of radiological protocols using machine learning. In *Proceedings of the 2011 workshop on Data mining for medicine and healthcare* (pp. 52-55). ACM.

[186] Cole, T. S., Frankovich, J., Iyer, S., LePendu, P., Bauer-Mehren, A., & Shah, N. H. (2013). Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: a new model for EHR-based research. *Pediatric Rheumatology*, *11*(1), 45.

[187] Warrer, P., Hansen, E. H., Juhl-Jensen, L., & Aagaard, L. (2012). Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *British journal of clinical pharmacology*, *73*(5), 674-684.

[188] Luo, Z., Miotto, R., & Weng, C. (2013). A human–computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. *Journal of biomedical informatics*, *46*(1), 33-39.

[189] Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, *121*, 279.

[190] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

[191] Pivovarov, R., & Elhadad, N. (2015). Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, *22*(5), 938-947.

[192] Elhadad, N., Kan, M. Y., Lok, S., & Muresan, S. (2001, January). PERSIVAL: personalized summarization over multimedia health-care information. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries* (p. 455). ACM.

[193] Fiszman, M., Rindflesch, T. C., & Kilicoglu, H. (2004, May). Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT-NAACL workshop on computational lexical semantics* (pp. 76-83). Association for Computational Linguistics.

[194] Arnold, P., & Rahm, E. (2015). SemRep: A repository for semantic mapping. *Datenbanksysteme für Business, Technologie und Web (BTW 2015)*.

[195] Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, *32*(suppl_1), D267-D270.

[196] Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. I. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(suppl_1), i180-i182.

[197] Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., & Rogers, W. J. (2004). The NLM indexing initiative's medical text indexer. Medinfo, 89.

[198] Spyropoulos, C. D., & Karkaletsis, V. (2005). Information extraction and summarization from medical documents.

[199] Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from medical documents: a survey. *Artificial intelligence in medicine*, *33*(2), 157-177.

[200] Rindflesch, T. C., Kilicoglu, H., Fiszman, M., Rosemblat, G., & Shin, D. (2011). Semantic MEDLINE: An advanced information management application for biomedicine. *Information Services & Use*, *31*(1-2), 15-21.

[201] Workman, T. E., & Stoddart, J. M. (2012). Rethinking information delivery: using a natural language processing application for point-of-care data discovery. *Journal of the Medical Library Association: JMLA*, *100*(2), 113.

[202] Yepes, A. J. J., Plaza, L., Carrillo-de-Albornoz, J., Mork, J. G., & Aronson, A. R. (2015). Feature engineering for MEDLINE citation categorization with MeSH. BMC bioinformatics, 16(1), 113.

[203] Zhang, H., Fiszman, M., Shin, D., Miller, C. M., Rosemblat, G., & Rindflesch, T. C. (2011). Degree centrality for semantic abstraction summarization of therapeutic studies. *Journal of biomedical informatics*, *44*(5), 830-838.

[204] Moral, C., de Antonio, A., Imbert, R., & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval. *Information Research: An International Electronic Journal*, *19*(1), n1.

[205] Sarkar, K., Nasipuri, M., & Ghose, S. (2011). Using machine learning for medical document summarization. *International Journal of Database Theory and Application*, *4*(1), 31-48.

[206] Danforth, K. N., Early, M. I., Ngan, S., Kosco, A. E., Zheng, C., & Gould, M. K. (2012). Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. *Journal of Thoracic Oncology*, *7*(8), 1257-1262.

[207] Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, *42*(5), 950-966.

[208] Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., ... & Wheeldin, B. (2007). The CLEF corpus: semantic annotation of clinical text. In *AMIA Annual Symposium Proceedings* (Vol. 2007, p. 625). American Medical Informatics Association.

[209] Bontcheva, K., Cunningham, H., Tablan, V., Maynard, D., & Hamza, O. (2002, July). Using GATE as an Environment for Teaching NLP. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1* (pp. 54-62). Association for Computational Linguistics.

[210] Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, *47*, 1-10.

[211] Fabian, G., Wächter, T., & Schroeder, M. (2012). Extending ontologies by finding siblings using set expansion techniques. *Bioinformatics*, *28*(12), i292-i300.

[212] Luther, S., Berndt, D., Finch, D., Richardson, M., Hickling, E., & Hickam, D. (2011). Using statistical text mining to supplement the development of an ontology. *Journal of Biomedical Informatics*, *44*, S86-S93.

[213]Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology*, *46*(1), 561-584.

[214] Saldanha, G. (2009). Principles of corpus linguistics and their application to translation studies research. *Tradumàtica: traducció i tecnologies de la informació i la comunicació*, (7).

[215] Van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., ... & Furlong, L. I. (2012). The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, *45*(5), 879-884.

[216] Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A. D., & Casillas, A. (2015). On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of biomedical informatics*, *56*, 318-332.

[217] Vasant, D., Neff, F., Gormanns, P., Conte, N., Fritsche, A., Staiger, H., & Robinson, P. (2015). DIAB: an ontology of type 2 diabetes stages and associated phenotypes. *Proceedings of Phenotype Day at ISMB*, *2015*, 24-27.

[218] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, *7*(Jan), 1-30.

[219] Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011, September). On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 145-158). Springer, Berlin, Heidelberg.

[220] Reyes, O., Morell, C., & Ventura, S. (2015). Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context. *Neurocomputing*, *161*, 168-182.

[221] Charte, F., & Charte, D. (2015). Working with Multilabel Datasets in R: The mldr Package. *R Journal*, *7*(2).

[222] Fang, Y. C., Huang, H. C., Chen, H. H., & Juan, H. F. (2008). TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. *BMC complementary and alternative medicine*, *8*(1), 58.

[223] Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, *17*(6), 646-651.

[224] Metais, E., Nakache, D., & Timsit, J. F. (2006, May). Automatic classification of medical reports, the CIREA project. In *Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey* (pp. 354-359).

[225] Abelleira, M. A. P., & Cardoso, C. A. (2010). Minería de texto para la categorización automática de documentos. *PhD in Computer Science por Carnegie Mellon University, Madrid, España*.

[226] Lussier, Y. A., Shagina, L., & Friedman, C. (2001). Automating SNOMED coding using medical language understanding: a feasibility study. In Proceedings of the AMIA Symposium (p. 418). American Medical Informatics Association.

[227] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

[228] Cárdenas, J., Olivares, G., & Alfaro, R. (2014). Clasificación automática de textos usando redes de palabras. *Revista signos*, *47*(86), 346-364.

[229] Gibaja, E., & Ventura, S. (2015). A tutorial on multilabel learning. *ACM Computing Surveys (CSUR)*, *47*(3), 52.

[230] Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, *11*(5), 392-402.

[231] Gibaja, E., & Ventura, S. (2014). Multi-label learning: a review of the state of the art and ongoing research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(6), 411-444.

[232] Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, *26*(8), 1819-1837.

[233] Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, *18*, 1-25.

[234] Santos, A., Canuto, A., & Neto, A. F. (2011). A comparative analysis of classification methods to multi-label tasks in different application domains. *Int. J. Comput. Inform. Syst. Indust. Manag. Appl*, *3*, 218-227.

[235] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215-e220.

[236] Kalra, S., Li, L., & Tizhoosh, H. R. (2019). Automatic Classification of Pathology Reports using TF-IDF Features. *arXiv preprint arXiv:1903.07406*.

[237] Yuan, M., Ouyang, Y. X., & Xiong, Z. (2013). A text categorization method using extended vector space model by frequent term sets. *Journal of Information Science and Engineering*, *29*(1), 99-114.

[238] Keikha, M., Khonsari, A., & Oroumchian, F. (2009). Rich document representation and classification: An analysis. *Knowledge-Based Systems*, *22*(1), 67-71.

[239] De Maio, C., Fenza, G., Loia, V., & Parente, M. (2019). Text Mining Basics in Bioinformatics.

[240] Kim, W., Aronson, A. R., & Wilbur, W. J. (2001). Automatic MeSH term assignment and quality assessment. In *Proceedings of the AMIA Symposium* (p. 319). American Medical Informatics Association.

[241] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, *34*(5), 301-310.

[242] Rokach, L., Romano, R., & Maimon, O. (2008). Negation recognition in medical narrative reports. *Information Retrieval*, *11*(6), 499-538.

[243] McCray, A. T. (1989, November). The UMLS Semantic Network. In *Proceedings. Symposium on Computer Applications in Medical Care* (pp. 503-507). American Medical Informatics Association.

[244] Fine, S., Singer, Y., & Tishby, N. (1998). The hierarchical hidden Markov model: Analysis and applications. *Machine learning*, *32*(1), 41-62.

[245] Smith, L., Rindflesch, T., & Wilbur, W. J. (2004). MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, *20*(14), 2320-2321.

[246] Schuemie, M. J., Sen, E., 't Jong, G. W., van Soest, E. M., Sturkenboom, M. C., & Kors, J. A. (2012). Automating classification of free-text electronic health records for epidemiological studies. *Pharmacoepidemiology and drug safety*, *21*(6), 651-658.

[247] Luque, C., Luna, J. M., Luque, M., & Ventura, S. An advanced review on text mining in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1302.

[248] Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2.

[249] Białecki, A., Muir, R., Ingersoll, G., & Imagination, L. (2012, August). Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval* (p. 17).

[250] Lourenço, A., Carreira, R., Carneiro, S., Maia, P., Glez-Peña, D., Fdez-Riverola, F., ... & Rocha, M. (2009). @ Note: a workbench for biomedical text mining. *Journal of biomedical informatics*, *42*(4), 710-720.

[251] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... & Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, *11*(1), 26-31.

[252] Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., & Vlahavas, I. (2011). Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, *12*(Jul), 2411-2414.

[253] Read, J., Reutemann, P., Pfahringer, B., & Holmes, G. (2016). Meka: a multi-label/multi-target extension to weka. *The Journal of Machine Learning Research*, *17*(1), 667-671.

[254] Holmes, G., Donkin, A., & Witten, I. H. (1994, November). WEKA: a machine learning workbench. In *Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference* (pp. 357-361). IEEE.

[255] Hofmann, M., & Klinkenberg, R. (Eds.). (2013). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

[256] Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3-4), 327-348.

[257] Schapire, R. E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine learning*, *39*(2-3), 135-168.

[258] Névéol, A., Shooshan, S. E., Humphrey, S. M., Mork, J. G., & Aronson, A. R. (2009). A recent advance in the automatic indexing of the biomedical literature. *Journal of biomedical informatics*, *42*(5), 814-823.

[259] Rak, R., Kurgan, L., & Reformat, M. (2005, December). Multi-label associative classification of medical documents from medline. In *Fourth International Conference on Machine Learning and Applications (ICMLA'05)* (pp. 8-pp). IEEE.

[260] Botsis, T., Nguyen, M. D., Woo, E. J., Markatou, M., & Ball, R. (2011). Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *Journal of the American Medical Informatics Association*, *18*(5), 631-638.

[261] Cunningham, H., Gaizauskas, R. J., & Wilks, Y. (1995). *A general architecture for text engineering (GATE): A new approach to language engineering R & D*. University of Sheffield, Depar of Computer Science.

[262] Lakhani, P., Kim, W., & Langlotz, C. P. (2012). Automated detection of critical results in radiology reports. *Journal of digital imaging*, *25*(1), 30-36.

[263] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium* (p. 17). American Medical Informatics Association.

[264] Divoli, A., & Attwood, T. K. (2005). BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics*, *21*(9), 2138-2139.

[265] Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., & Tsujii, J. I. (2006, July). Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 1017-1024). Association for Computational Linguistics.

[266] Tudela, P., Mòdol, J. M., Rego, M. J., Bonet, M., Vilaseca, B., & Tor, J. (2005). Error diagnóstico en urgencias: relación con el motivo de consulta, mecanismos y trascendencia clínica. Medicina clínica, 125(10), 366-370.

[267] Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., & Wishart, D. S. (2008). PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic acids research*, *36*(suppl_2), W399-W405.

[268] Goldstein, I., Arzumtsyan, A., & Uzuner, Ö. (2007). Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. In AMIA Annual Symposium Proceedings (Vol. 2007, p. 279). American Medical Informatics Association.

[269] Karimi, S., Dai, X., Hassanzadeh, H., & Nguyen, A. (2017). Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. BioNLP 2017, 328-332.

[270] Raju, M. K., Subrahmanian, S. T., & Sivakumar, T. (2017). A comparative survey on different text categorization techniques. *International Journal of Computer Science and Engineering*, *5*(3), 1612-1618.

[271] Pereira, L., Rijo, R., Silva, C., & Agostinho, M. (2013). ICD9-based text mining approach to children epilepsy classification. Procedia Technology, 9, 1351-1360.

[272] Pakhomov, S. V., Buntrock, J. D., & Chute, C. G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association*, *13*(5), 516-525.

[273] Suominen, H., Ginter, F., Pyysalo, S., Airola, A., Pahikkala, T., Salanterä, S., & Salakoski, T. (2008, July). Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. In Proceedings of the ICML/UAI/COLT Workshop on Machine Learning for Health-Care Applications.

[274] Cohen, W. W. (1995). Fast effective rule induction. In *Machine Learning Proceedings 1995* (pp. 115-123). Morgan Kaufmann.

[275] Wang, Y., Yu, Z., Jiang, Y., Liu, Y., Chen, L., & Liu, Y. (2012). A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records. *Journal of Biomedical Informatics*, *45*(2), 210-223.

[276] Sharma, D. K., & Hota, H. S. (2013). Data mining techniques for prediction of different categories of dermatology diseases. *Journal of Management Information and Decision Sciences*, *16*(2), 103.

[277] Suzuki, T., Yokoi, H., Fujita, S., & Takabayashi, K. (2008). Automatic DPC code selection from electronic medical records. *Methods of information in medicine*, *47*(06), 541-548.